Volume 47 Number 1 March 2023 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Enhancement of NTSA Secure Communication with One-Time Pad (OTP) in IoT

Ali Hasan Aidaros Alattas[1], Mahmood A. Al-Shareeda[1], Selvakumar Manickam[1,*] and Murtaja Ali Saare[2]
[1]National Advanced IPv6 Centre, Universiti Sains Malaysia, 11800, Penang, Malaysia
[2]Department of Computer Technology Engineering, Shatt Al-Arab University College, Basrah, Iraq
E-mail: alshareeda022@usm.my, selva@usm.my, murtaja.a.sari@sa-uc.edu.iq
*Corresponding author

Keywords: Internet of Things (IoT), NTSA, One-Time Pad (OTP), lightweight cryptographic algorithms

Received: October 24, 2022

Internet of Things (IoT) systems use interconnected devices with limited processing, memory, storage, and power availability. Designing the IoT system requires careful consideration of data security. IoT networks are used to collect, process, and transport data; as a result, it needs to be encrypted and secured. To ensure that the data of IoT systems are protected, a variety of lightweight encryption techniques have been developed. These algorithms are unable to carry out complicated or extensive computations. The current challenge facing lightweight cryptographic algorithms, such as NTSA, is how to combine the highest level of security with the least amount of negative influence on runtime speed and space. By applying the One-Time Pad (OTP) technique, the proposed mechanism can raise the security level and effectiveness of NTSA. The proposed mechanism must be put into practise and put to the test in order to demonstrate its effectiveness and capacity to satisfy the needs of the resource-constrained devices. Due to the benefits of the OTP, this suggested method would be beneficial for devices with minimal resources. The proposed technique offers a greater security level, 2134, than NTSA, 2128, after examining and evaluating the experimental data noticed throughout the tests. NTSA is slower than the suggested approach by 70% in terms of runtime speed. While NTSA uses 16% of SRAM, the proposed algorithm only uses 12%. NTSA uses 70% more energy than the suggested algorithm, with higher energy consumption results of 0.000388 Joules for the proposed algorithm and 0.001295 Joules for NTSA.

Povzetek: Predstavljena je nova metoda za šifriranje in varnostna vprašanja IoT omrežij, ki dosega boljše rezultate kot NTSA.

1 Introduction

The Internet is a system architecture that has allowed communications to advance to connect devices via different networks all over the world. Any individual object that connects to one of its networks can access the Internet for nearly any purpose that requires information (1; 2). It enables access to digital information through human or machine-to-machine (M2M) communications (3). Each connected object in the Internet of Things has a unique identity and can connect to other connected objects (4). Medical equipment, monitoring equipment, machinery, automobiles, and buildings will all be upgraded to become intelligent objects that can interact with people or other IoT devices (5; 6). The digital transformation of many industries is what fuels the IoT's growth. IoT connections will increase from fifteen billion in 2015 to seventy-five billion by 2025, as stated in (7), see Figure 1.

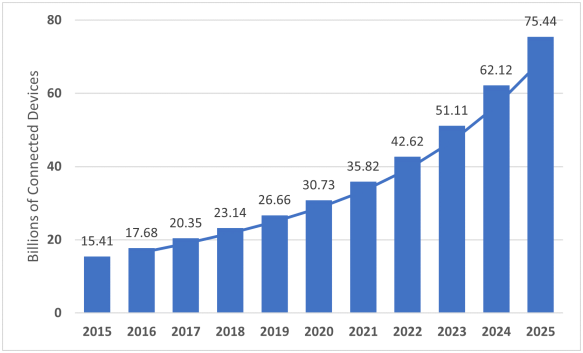


Figure 1: IoT connected devices in number.

The security issue is an afterthought because the resource-constrained networked device is meant to consume a little power to give all essential capabilities (8; 9). There are problems with IoT hardware, including the possibility of an attack on the device's encrypted data since some IoT devices are too small to support asymmetric cryptography algorithms. A gadget transmits or receives data that needs to be encrypted (10; 11). However, using cryptographic methods on devices with limited resources is difficult. The device itself, such as an 8-bit microcontroller with a 2KB RAM limit, performs the encryption operation (12; 13).

Traditional cryptography techniques cannot be implemented on such devices since they are expensive and inefficient. In order to address the security concerns on nodes with limited resources, lightweight ciphers have been developed. They are made to achieve cryptographic computational operation while adhering to the restrictions of microcontrollers, small-size RAM, and low power consumption (14; 15; 16).

By exploiting the benefits of the OTP technique, the suggested mechanism introduces a solution for both high security and higher performance. This paper focuses on symmetric encryption ciphers and the OTP approach as a foundation for lightweight cryptography. The advantages of block ciphers, which are simple to implement, the OTP technique, high security, and high performance can result in a dependable and robust system. The following are some of the research's contributions:

- This study will make it easier to deploy OTP in all areas of life and execute its secrecy into sensitive applications that require a high level of security with high performance because the OTP approach has shortcomings that have limited its adoption.
- One of OTP's flaws is the key exchange procedure. Therefore, this study will address this problem by creating a simple protocol for parties to exchange keys.

The rest of this paper is organized as follows. Section 2 reviews some related work. Section 3 introduces the background of this paper. Section 4 describes the general proposed mechanism's architecture. Section 5 and Section 6 provide a security analysis and results for the proposed mechanism, respectively. Lastly, Section 7 shows the conclusion and future work in this work.

2 Related work

Advance Encryption Standard (AES) was proven to be the best trusted and researched block cipher and still has to be subjected to more study to make it acceptable for resource-constrained devices, as indicated in (17; 18). While some lightweight cryptographic algorithms, like G-TBSA, are adequate for some factors like processing power and energy, they are not resistant to all types of attacks. Like G-TBSA, a number of lightweight cryptographic algorithms are adequate in some respects, such as computational power and energy, but are not resistant to other assaults.

None of the prominent modern lightweight block and stream ciphers is typical in offering the security, affordability, and performance for IoT devices with limited resources (19). It has been noted that the advancement of lightweight cryptography is still ongoing (20; 21).

However, developing an algorithm that satisfies the needs of lightweight cryptography for IoT devices is a considerable task. To accommodate various IoT device memory limits, the author has devised a simple encryption method that employs variable-sized keys and data blocks. This idea makes use of DNA sequences to produce random keys.

Many current LWC algorithms, according to (22), concentrate on lowering the cost of memory, computational power, physical area, and energy consumption and enhancing throughput and latency without paying attention to security vulnerabilities. In addition, the author claims that a successful encryption algorithm must strike a balance between three LWC design objectives (Security, Performance, Cost).

Banani et al. (23) employed the standard performance measures (memory occupied, execution time, and power consumption) to trade off among the various algorithms, including TEA. They did this by referring to the security and performance evaluation criteria. The avalanche effect attribute was utilised by the author to illustrate the security metrics in (24).

In order to summary the limitations of existing works, we list algorithms and attacks occurred as presented in Table 1. According to this leak, we enhance NTSA secure communication with OTP in IoT in order to raise the security level and effectiveness of NTSA. The proposed mechanism must be put into practise and put to the test in order to demonstrate its effectiveness and capacity to satisfy the needs of the resource-constrained devices (13).

3 Background

3.1 One-Time Pad technique

Similar to a stream cipher but not one that uses a random key generator is a One-Time Pad (OTP). It is a safe method for encrypting a message so that a cryptanalyst cannot decipher the message from the information (25). When encrypting and decrypting data, a random key must have a length equal to or greater than the message length produced by a genuine random generator. Then, it will be deleted so that a fresh new random key is used for the subsequent encryption and decryption procedures (26; 27).

OTP typically employs the XOR operation to encrypt plaintext by fusing the message and key bits, which is quick and appropriate for IoT devices. This increases security and makes OTP uncrackable under the following circumstances: (I) The key's unpredictability; (II) The length of the key must be at least as long as the plaintext; (IV) The key can be used just once; and (IV) The key has a very high level of confidentiality (28; 29; 30).

3.2 Lightweight Cryptographic Algorithm (LWC)

Designed for devices with limited resources, Lightweight Cryptographic Algorithm (LWC) is a branch of cryptography that seeks to offer solutions (31). The NIST started a lightweight cryptography project in 2013 to investigate how well the NIST-approved cryptographic standards function on restricted devices and to determine the demand for

Table 1: Different Attacks on Some Lightweight Cryptosystems in Related Work

Algorithm	Attack	Cipher	Key Size	Structure
TEA	Related-key attack	Block	128 bits	Feistel
XTEA	Related-key attack	Block	128 bits	Feistel
HB-2	Related-key attack	Hybrid	128 bits	Hybrid
PRINTcipher	Related-key attack	Block	80, 160 bits	SPN
PRESENT	Related-key attack	Block	80, 128 bits	SPN
XXTEA	Chosen-Plaintext attack	Block	128 bits	Feistel
AES	Biclique cryptanalysis	Block	128, 192, 256 bits	SPN
LED	Biclique cryptanalysis	Block	64, 80, 96, 128 bits	SPN
PRESENT	Biclique cryptanalysis	Block	80, 128 bits	SPN
Grain	Key recovery attack	Stream	80 bit	Stream
MICKEY	Differential fault attack	Stream	80 bits	Stream
SIMON	Differential fault attack	Block	64,72, 96,128, 144, 192, 256 bits	Feistel
SPECK	Differential fault attack	Block	64,72, 96,128, 144, 192, 256 bits	ARX
PRESENT	Differential fault attack	Block	80, 128 bits	SPN
PRESENT	Truncated differential attack	Block	80, 128 bits	SPN
ChaCha	Truncated differential attack	Stream	256 bits	ARX

specific lightweight cryptography standards. The literature will go into detail about how lightweight encryption algorithms have been designed to meet the capabilities of resource-constrained devices to provide both a high level of security and high performance in terms of minimizing the runtime and space complexities as much as feasible (32).

3.3 TEA and NTSA algorithms

TEA uses 64 rounds spread over 32 cycles. Starting with dividing a 128-bit key into four 32-bit subkeys (k0, k1, k2, and k3), a 128-bit plaintext block is split into two blocks of 32 bits. Each set of four operations uses ADD, XOR, and left and right shift operations. In order to increase confusion during all rounds of encrypting a 64-bit plaintext block, NTSA, which is an upgrade to TEA, tries to generate dynamically changing subkeys derived from a 128-bit key (33; 34).

4 General proposed mechanism's architecture

The TEA algorithm performs well in LWC and is simple to implement in both hardware and software. It also uses less memory. However, it is susceptible to related-key assaults and has a flaw in the round function mixing. Based on the findings of the comparison analysis between TEA and NTSA, NTSA resolved the primary scheduling issue (35). It turns out that developing a system based on the NTSA and OTP will offer a reliable and lightweight cryptosystem for IoT devices with limited resources. As shown in Figure 2, the system uses block ciphers and OTP techniques along with two different forms of symmetric-key primitives (36).

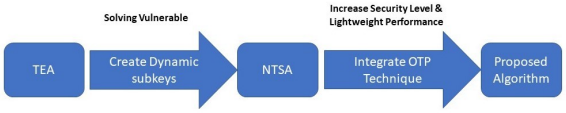


Figure 2: The proposed scheme's mechanism.

4.1 Random keys generation

Both in hardware and software, the TEA algorithm works well in LWC and is straightforward to implement. It consumes less memory as well. However, it has a vulnerability in the round function mixing and is vulnerable to related-key attacks. The fundamental scheduling issue was resolved by NTSA based on the results of the comparative analysis between TEA and NTSA (12).

To prevent noise bias between the axes, the Von Neumann extractor method extracts two bits from each axis. Then the desired value is generated by applying Equation (1) to a random byte.

$$RandByte = (x \leq 6) \oplus (z \leq 4) \oplus (y \leq 2) \oplus x \oplus (z \leq 2) \quad (1)$$

Additionally, XORing independent binary variables always minimizes bias, as the piling-up lemma in (37) shows. Let the random byte be made up of the values x, y, and z that are retrieved from the x, y, and z axes, respectively. The flowchart for the procedure that will produce a random byte is shown in Figure 3.

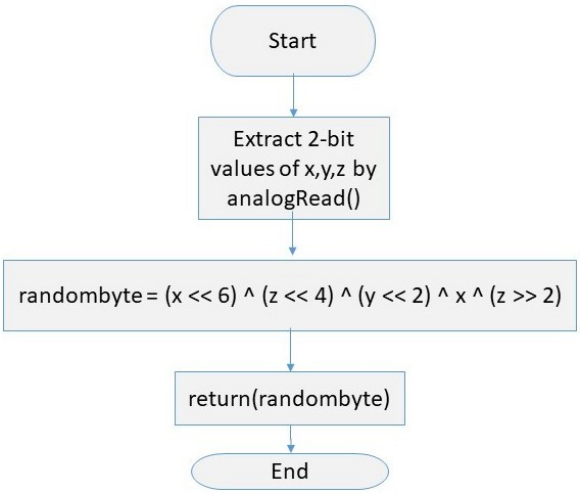


Figure 3: RNG flowchart.

4.2 Proposed mechanism

We examine three different instances of data transmission via the Internet of Things devices:

- Periodically, devices will send data; the transmission interval is determined by the application domain.
- The transmission interval is fixed, and devices will periodically deliver data.
- Data will be sent by devices when it is modified.

This approach is intended to be used in the third scenario, which involves passing information from the temperature sensor to an air conditioner. A key must be used once in the OTP approach before being discarded in order to generate a new key for use in the following encryption procedure. Therefore, sending data on a regular basis is not recommended, especially if the interval is small, like every second or even every hour (16).

4.2.1 Encryption algorithm

Regarding the first research issue, Figure 4 provides an illustration of the suggested algorithm. The Feistel structure, which uses the around function, is used by the proposed method since it employs the same round function that NTSA and TEA do. Data block P and subkey ki are two inputs that a round function accepts and returns one result.

The following conditions can be met by watching the encryption process's algorithm:

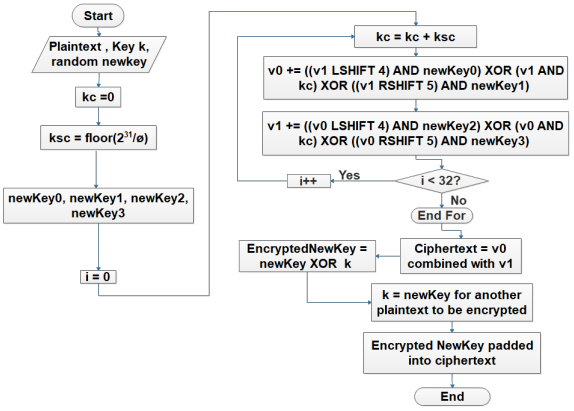


Figure 4: Encryption flowchart.

- single-use key.
- The key's confidentiality.
- True random secret keys for each encryption procedure and the high level of security provided by random keys contribute to increasing security.

4.2.2 Decryption algorithm

Figure 5 shows depicts the entire decryption process.

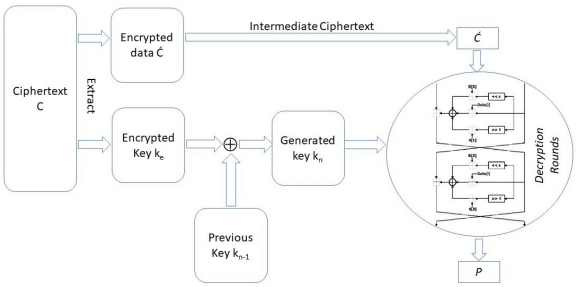


Figure 5: The entire decryption process.

4.3 Key padding protocol

When a new key is generated and utilised in each encryption procedure, the problem of key exchange between a sender and recipient arises. The sender and the recipient need to exchange this key. The problem must be taken into account to minimise the need for computationally intensive operations, as is the case with traditional cryptography like RSA. The complexity of key exchange problems grows as a result of the connectivity between IoT devices and machine-to-machine communication. The approach suggests padding the key for message encryption and decryption into the ciphertext after it has been encrypted using the previous key and the XOR operation, as shown in Figure 6.

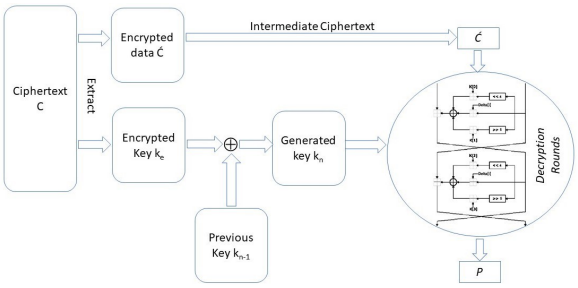


Figure 6: Process of encrypted key padding.

4.4 Key extraction protocol

Using (16), the intermediate ciphertext's bytes are first extracted from the final ciphertext C during the decryption process, D(cn, kn-1). Next, the bytes of the encrypted key ke is collected from the final ciphertext, as shown in Figure 7. Then kn is obtained by performing an XOR operation between kn-1 and ke in order to decrypt the message and the following newKey ke.

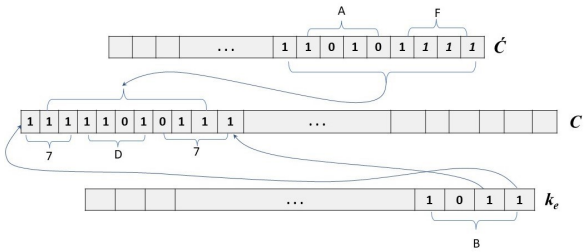


Figure 7: Final ciphertext bits.

4.5 Discussion

OTP is the suggested remedy, as discussed in the sections that came before it, to simplify the design complexity of lightweight encryption methods. If all of its requirements are met, it operates at a high-performance level with strong security. NTSA was chosen for this thesis because it fixed a flaw in the TEA algorithm, the most desirable lightweight encryption technique. However, NTSA continues to employ the same key, which is vulnerable to attack, across all encryption processes. IoT systems' limited-resource devices can use the proposed technique. The simplest and fastest computational processes, XOR, left and right shift, and modular addition arithmetic, which rely on bitwise operation, have been suggested as an effective mechanisms for implementing key exchange procedures.

5 Security analysis

Shift registers, Feistel structures, and substitution-permutation networks are a few examples of specific structures on which certain ciphers are based. The most frequent threats to Feistel-structure block ciphers, upon which this paper depends, will be covered.

5.0.1 Cipher-text only attack

In this type of attack, the attacker can capture ciphertext and attempt to decrypt it in order to learn more about the plaintext and, if possible, the key. To examine and decrypt ciphertexts, an attacker needs n of them. These attacks have not been successful against current ciphers.

5.0.2 Known-plaintext attack

Some ciphertext's plaintext can be deciphered by an attacker. The goal of this assault is to reveal and decrypt the remaining ciphertext blocks using already-known information, which may reveal the key.

5.0.3 Chosen-plaintext attack

Despite being the least factual, this form of attack is potent. With this technique, the key used to encrypt data is determined by measuring a change in the ciphertext.

5.0.4 Chosen-ciphertext attack

The chosen-Ciphertext attack also includes a chosen-plaintext assault, which decrypts ciphertexts with a particular key. If this type of attack is combined with a chosen-plaintext attack, it is not very practical.

5.0.5 Differential cryptanalysis

The typical technique for attacking cryptographic algorithms is this one. Since linear cryptanalysis uses a known-plaintext attack instead of the usual differential cryptanalysis method's chosen-plaintext attack, it is thought to be more practical in everyday life. Particularly, it examines ciphertext pairs. Pairs of ciphertexts with distinct plaintext differences and examine how these differences change as the plaintexts move through the encryption algorithm's rounds when they are encrypted with the same key. As long as the two plaintexts satisfy specific differences, they can be selected at random (with a fixed difference). Then, assign various probabilities to various keys based on the variations in the generated ciphertexts. One key will become more and more obvious as the most likely correct key as more and more ciphertexts are studied.

5.0.6 Related-Key attack

Comparable to differential cryptanalysis, but focused on key differences. Without knowing the actual keys, this approach focuses on the relationship between a pair of keys. It uses plaintext encryption using both the real key K and some derived keys, as well as a straightforward link between subkeys in neighboring rounds. The method for

changing the keys must be specified; it may involve flipping key bits while concealing the true key.

The TEA's issue, which is brought on by weak key scheduling and a weak mixing component of the round function, is resolved by the NTSA. The TEA technique can be broken by a related-key attack using 223 selected plaintexts, especially if the key is weak (38). The NTSA's defence against the related-key attack will then be clear see how it was created.

5.0.7 Keys equivalence

If two keys, k1, and k2, produce the same ciphertext after encrypting the same plaintext, then they are equal. Ek1(P) = Ek2(P), where E is the encryption function, P is the plaintext, and k1 and k2 are separate keys K and K is the key space, illustrating this relationship. The connection between the classes that make up K is such that k1 and k2 are members of the same class. To make this argument more understandable, use mathematical equations and demonstrate the TEA's susceptibility as shown in (39).

5.0.8 Resistance of NTSA against related-key attack

The NTSA uses the same round function as the TEA algorithm, with one modification to improve the key schedule procedure. In the NTSA, the extract() function is called after each round, and it dynamically returns a value from an array. Thirty-three separate 32-bit values that are obtained from the 128-bit key fill up this array

5.0.9 The proposed mechanism security analysis against related-key attack

Great security level and high performance in terms of space and time complexity are coupled in the suggested method by incorporating the OTP technique. As long as its requirements are met, the OTP, a conventional but nonetheless powerful cipher, can withstand quantum computers (40). Despite using the same round function as the TEA, the suggested approach is more secure than NTSA.

6 Results

6.1 Execution time

This analysis will show the encryption and decryption process execution times for both algorithms, measured in milliseconds based on the number of cycles. The number of bits encrypted and decrypted using the 128-bit key serves as a measure of the data size, which is determined by each cycle's two rounds. Tables and bar charts are going to be used to show the results. With data blocks of 64, 128, 192, and 256 bits, encryption and decryption functions will be conducted throughout the number of cycles 8, 16, and 32 to reach the execution time tests. These several categories serve as illustrations of how the suggested mechanism and the performance of NTSA are affected by the number of rounds and size of the data block.

6.1.1 Execution time of encryption process

The encryption function in the suggested technique requires three inputs: a 64-bit block of data, a previous key with a 128-bit size, and a 128-bit fresh key that is generated before each new encryption function begins. It has two parameters in NTSA: plaintext block with a 128-bit key and 64 bits. The execution time of NTSA increases by roughly 0.828 ms in Figure 8 and Table 2 for the same number of rounds, 16 rounds, and various block sizes. The suggested algorithm, however, is implemented more quickly than NTSA, and its runtime increases by about 0.544 ms for every increase in block size. In other words, the suggested algorithm outperforms the NTSA by 50%.

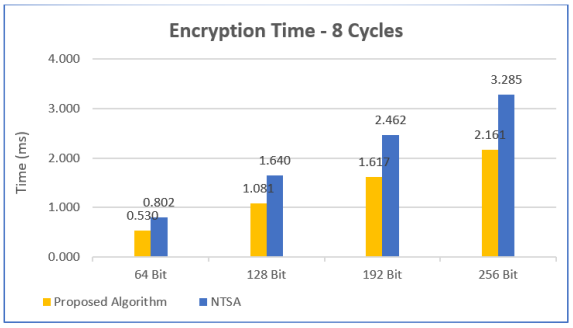


Figure 8: Encryption time for 8 cycles in ms.

Table 2: 8-cycle encryption process results in milliseconds

Algorithm	64 Bits	128 Bits	192 Bits	256 Bits
NTSA	0.802	1.640	2.462	3.285
Proposed	0.530	1.081	1.617	2.161

6.2 Execution time of decryption process

While the proposed algorithm's ciphertext contains both the encrypted data and the new key, the NTSA's decryption function requires 64-bit ciphertext and 128-bit key parameters. The execution time increment rate for both algorithms to complete the decryption function in 8 cycles, or 16 rounds, is shown in Figure 9 and Table 3. The proposed technique and the NTSA have slightly different runtimes for the encryption and decryption operations under a class of eight cycles.

6.3 Memory occupation

Memory occupation in Bytes: In this paper, memory usage is calculated using SRAM memory for execution time and

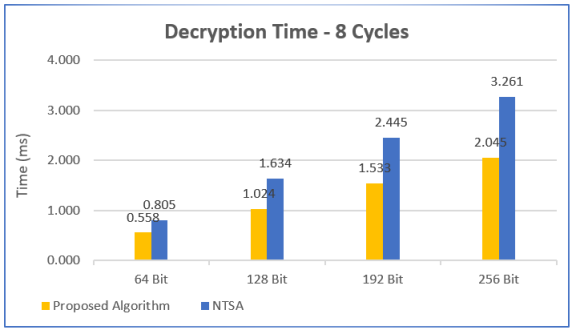


Figure 9: Decryption time for 8 cycles in ms.

Table 3: 8-cycle decryption process results in milliseconds

Algorithm	64 Bits	128 Bits	192 Bits	256 Bits
NTSA	0.805	1.634	2.445	3.261
Proposed	0.558	1.024	1.533	2.045

flash memory for storing code. To measure the amount of energy used by both algorithms, the Arduino Uno board is powered by a 9-volt battery in this experiment. The reading of the current passing through the Arduino Uno board was taken using the multimeter. There is a 5V voltage and a 20mA current (0.02A).

Figures 10 and 11 show that the NTSA uses 7% of flash memory to store the algorithm, which is 2546B, and 16%, or 340 bytes of 2KB, to store the global variables in SRAM for encrypting and decrypting 64-bit plaintext with a 128-bit key. Because the NTSA's code file has two routines—encryption and decryption—as well as one function to retrieve the array's contents, it uses less flash memory than the suggested approach. In contrast, the code file for the proposed approach contains the encryption and decryption procedures as well as the key generation function. In comparison, the NTSA employs an array to hold 33 32-bit subkeys during runtime, which requires more SRAM capacity. The suggested approach, in contrast, employs an array that holds six 32-bit values that constitute the final ciphertext.

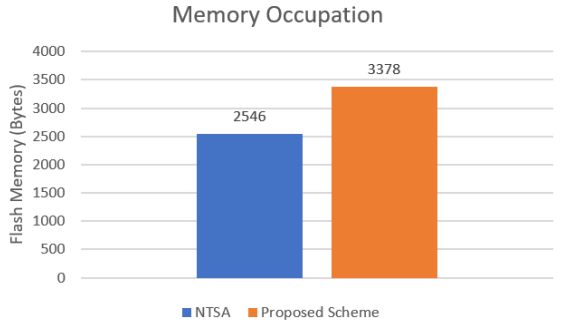


Figure 10: Memory occupations.

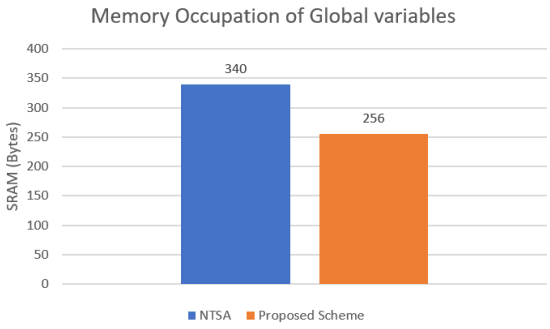


Figure 11: Global variable memory occupation.

6.4 Energy consumption

Energy consumption: a device with limited resources and low energy usage lasts longer on its battery. As shown in Figure 12, the following equipment should be available to conduct this experiment and assess the power consumption: an 8-bit microcontroller Arduino Uno board (MCU), the proposed algorithm, and its equivalent NTSA. (1) Multimeter to measure the voltage and the current; (2) Jumper wires; (3) Banana Plug to Crocodile Clip; (4) DC Barrel Jack Adapter – Female to screw terminals; and (5) Power Source whether 9V Battery with 9V Battery Connector to DC Jack Arduino or Wall Power Supply (5V- 2Am).

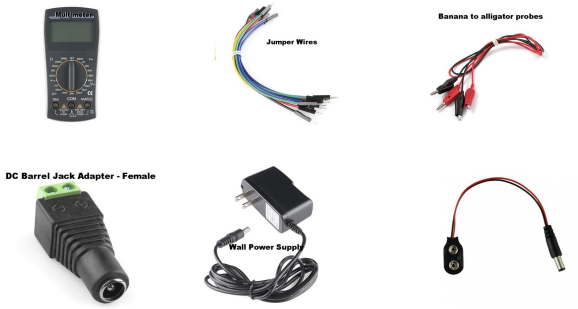


Figure 12: Tools used.

The procedures that follow explain how to set up the necessary equipment to begin this experiment's mechanism:

- Prepare the multimeter by inserting the red probe of the banana-crocodile cable into the mAV port to measure the voltage and the black probe of the cable into the COM port to measure the current.
- The multimeter's dial should be set to A for current and V for voltage.
- Connect the red probe from the multimeter to the (+) end of the power supply and the black probe to the Vin port on the Arduino Uno board using a DC Barrell Jack Adapter. Lastly, attach the (-) end of the power

supply to the GND port on the Arduino Uno board. This circuit has a series connection.

In this experiment, the Arduino Uno board is powered by a 9-volt battery to measure the energy required by both algorithms. Using the multimeter, the reading of the current flowing through the Arduino Uno board was captured. The voltage is 5V, and the current is equal to 20mA (0.02A). Table 4 shows the energy usage for the encryption procedure for various categories of data sizes with fewer than 64 rounds. The results in Table 4 demonstrate that the suggested method provides great optimization in terms of power usage compared to the NTSA.

Table 4: The energy consumption for encryption process

Algorithm	64 Bits	128 Bits	192 Bits	256 Bits
NTSA	0.0003192	0.00065	0.000972	0.001295
Proposed	0.000096	0.000193	0.00029	0.000388

7 Conclusion and future work

Traditional cryptographic algorithms are not suitable for IoT devices due to their inherent limitations in terms of processing power, memory, storage, and energy. However, the ongoing development of lightweight cryptography will continue to produce suitable lightweight cryptographic mechanisms to meet these requirements. Consequently, this research suggests a mechanism to incorporate the OTP technique into the NTSA in order to take advantage of the high-security level with the high performance offered by the OTP and easy implementation offered by block cipher and combine them into one mechanism to provide a lightweight cryptographic algorithm that can be implemented on IoT devices easily and effectively.

The first research goal was accomplished by integrating the OTP technique into NTSA in order to increase security. The data was encrypted using various new random keys generated by the MPU6050 sensor, and the final ciphertext was created by padding the bits in order to share the newly generated key. The experiments covered in Chapter 4 demonstrate that the suggested mechanism offers a greater security level and higher performance in terms of the complexity of speed, reduced memory utilization, and lower energy consumption. This is relevant to the second study objective. The encryption and decryption runtimes show that NTSA is 70% slower than the suggested technique. NTSA uses 16% of SRAM, compared to 12% for the suggested method. In terms of security, the proposed technique offers 2134 security complexity compared to 2128 security complexity offered by NTSA. The proposed algorithm uses 0.000388 Joules of energy, but NTSA uses 0.0013 Joules, meaning that NTSA uses 70% more energy than the proposed approach.

References

[1] S. S. Oyewobi, K. Djouani, and A. M. Kurien, "Visible light communications for internet of things: Prospects and approaches, challenges, solutions and future directions," *Technologies*, vol. 10, no. 1, p. 28, 2022. [Online]. Available: https://doi.org/10.3390/technologies10010028

[2] M. A. Al-Shareeda, M. Anbar, I. H. Hasbullah, and S. Manickam, "Survey of authentication and privacy schemes in vehicular ad hoc networks," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 2422–2433, 2020. [Online]. Available: https://doi.org/10.1109/JSEN.2020.3021731

[3] I. H. Sarker, A. I. Khan, Y. B. Abushark, and F. Alsolami, "Internet of things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions," *Mobile Networks and Applications*, pp. 1–17, 2022. [Online]. Available: https://doi.org/10.1007/s11036-022-01937-3

[4] A. E. Omolara, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, H. Arshad *et al.*, "The internet of things security: A survey encompassing unexplored areas and new insights," *Computers & Security*, vol. 112, p. 102494, 2022. [Online]. Available: https://doi.org/10.1016/j.cose.2021.102494

[5] I. Ashraf, Y. Park, S. Hur, S. W. Kim, R. Alroobaea, Y. B. Zikria, and S. Nosheen, "A survey on cyber security threats in iot-enabled maritime industry," *IEEE Transactions on Intelligent Transportation Systems*, 2022. [Online]. Available: https://doi.org/10.1109/TITS.2022.3164678

[6] M. A. Al-Shareeda, M. Anbar, S. Manickam, and A. A. Yassin, "Vppcs: Vanet-based privacy-preserving communication scheme," *IEEE Access*, vol. 8, pp. 150 914–150 928, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3017018

[7] T. Alam, "A reliable communication framework and its use in internet of things (iot)," *CSEIT1835111| Received*, vol. 10, pp. 450–456, 2018. [Online]. Available: https://doi.org/10.36227/TECHRXIV.12657158.V1

[8] M. A. Al-shareeda, M. Anbar, S. Manickam, I. H. Hasbullah, N. Abdullah, M. M. Hamdi, and A. S. Al-Hiti, "Ne-cppa: A new and efficient conditional privacy-preserving authentication scheme for vehicular ad hoc networks (vanets)," *Appl. Math*, vol. 14, no. 6, pp. 1–10, 2020. [Online]. Available: https://doi.org/10.3390/s21248206

[9] M. Salimitari, M. Chatterjee, and Y. P. Fallah, "A survey on consensus methods in blockchain

for resource-constrained iot networks," *Internet of Things*, vol. 11, p. 100212, 2020. [Online]. Available: https://doi.org/10.36227/techrxiv.12152142

[10] S. Misra, A. Mukherjee, A. Roy, N. Saurabh, Y. Rahulamathavan, and M. Rajarajan, "Blockchain at the edge: Performance of resource-constrained iot networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 174–183, 2020. [Online]. Available: https://doi.org/10.1109/TPDS.2020.3013892

[11] V. Tambe, G. Bansod, S. Khurana, and S. Khandekar, "Reliability and availability of iot devices in resource constrained environments," *International Journal of Quality & Reliability Management*, 2022. [Online]. Available: https://doi.org/10.1108/IJQRM-09-2021-0334

[12] M. A. Al-Shareeda, M. Anbar, S. Manickam, and I. H. Hasbullah, "Password-guessing attack-aware authentication scheme based on chinese remainder theorem for 5g-enabled vehicular networks," *Applied Sciences*, vol. 12, no. 3, p. 1383, 2022. [Online]. Available: https://doi.org/10.3390/app12031383

[13] M. A. Al-shareeda, M. A. Alazzawi, M. Anbar, S. Manickam, and A. K. Al-Ani, "A comprehensive survey on vehicular ad hoc networks (vanets)," in *2021 International Conference on Advanced Computer Applications (ACA)*. IEEE, 2021, pp. 156–160. [Online]. Available: http://doi.org/10.1109/ACA52198.2021.9626779

[14] P. Panahi, C. Bayılmış, U. Çavuşoğlu, and S. Kaçar, "Performance evaluation of lightweight encryption algorithms for iot-based applications," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 4015–4037, 2021. [Online]. Available: https://doi.org/10.1007/s13369-021-05358-4

[15] M. A. Al-Shareeda, M. Anbar, I. H. Hasbullah, S. Manickam, and S. M. Hanshi, "Efficient conditional privacy preservation with mutual authentication in vehicular ad hoc networks," *IEEE Access*, vol. 8, pp. 144 957–144 968, 2020. [Online]. Available: https://doi.org/10.3390/su14169961

[16] M. A. Al-Shareeda, S. Manickam, M. A. Saare, and N. C. Arjuman, "Proposed security mechanism for preventing fake router advertisement attack in ipv6 link-local network," *Indones. J. Electr. Eng. Comput. Sci*, vol. 2023, no. 29, pp. 518–526, 2023. [Online]. Available: https://doi.org/10.11591/ijeecs.v29.i1.pp518-526

[17] I. K. Dutta, B. Ghosh, and M. Bayoumi, "Lightweight cryptography for internet of insecure things: A survey," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2019, pp. 0475–0481. [Online]. Available: https://doi.org/10.1109/CCWC.2019.8666557

[18] M. A. Al-Shareeda, S. Manickam, B. A. Mohammed, Z. G. Al-Mekhlafi, A. Qtaish, A. J. Alzahrani, G. Alshammari, A. A. Sallam, and K. Almekhlafi, "Provably secure with efficient data sharing scheme for fifth-generation (5g)-enabled vehicular networks without road-side unit (rsu)," *Sustainability*, vol. 14, no. 16, p. 9961, 2022. [Online]. Available: https://doi.org/10.3390/su14169961

[19] M. Rana, Q. Mamun, and R. Islam, "Current lightweight cryptography in iot security: A survey," in *Extended Abstracts*. Charles Sturt University, 2020, p. 27. [Online]. Available: https://researchoutput.csu.edu.au/ws/portalfiles/portal/100690557/SCM_HDR_Booklet_2020.pdf#page=27

[20] M. A. F. Al-Husainy, B. Al-Shargabi, and S. Aljawarneh, "Lightweight cryptography system for iot devices using dna," *Computers and Electrical Engineering*, vol. 95, p. 107418, 2021. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2021.107418

[21] M. A. Al-Shareeda and S. Manickam, "Man-in-the-middle attacks in mobile ad hoc networks (manets): Analysis and evaluation," *Symmetry*, vol. 14, no. 8, p. 1543, 2022. [Online]. Available: https://doi.org/10.3390/sym14081543

[22] V. A. Thakor, M. A. Razzaque, and M. R. Khandaker, "Lightweight cryptography algorithms for resource-constrained iot devices: A review, comparison and research opportunities," *IEEE Access*, vol. 9, pp. 28 177–28 193, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3052867

[23] S. Banani, S. Thiemjarus, K. Wongthavarawat, and N. Ounanong, "A dynamic light-weight symmetric encryption algorithm for secure data transmission via ble beacons," *Journal of Sensor and Actuator Networks*, vol. 11, no. 1, p. 2, 2021. [Online]. Available: https://doi.org/10.3390/jsan11010002

[24] W. Diaztary, D. Atmajaya, F. Umar, S. M. Abdullah *et al.*, "Tiny encryption algorithm on discrete cosine transform watermarking," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*. IEEE, 2021, pp. 415–420. [Online]. Available: https://doi.org/10.1109/EIConCIT50028.2021.9431930

[25] F. Ramadhani, U. Ramadhani, and L. Basit, "Combination of hybrid cryptography in one time pad (otp) algorithm and keyed-hash message authentication code (hmac) in securing the whatsapp communication application," *Journal of Computer Science, Information Technology and Telecommunication Engineering*, vol. 1, no. 1, pp. 31–36, 2020. [Online]. Available: https://doi.org/10.30596/jcositte.v1i1.4359

[26] A. Sarkar, S. R. Chatterjee, and M. Chakraborty, "Role of cryptography in network security," in *The" Essence" of Network Security: An End-to-End Panorama*. Springer, 2021, pp. 103–143. [Online]. Available: https://doi.org/10.1007/978-981-15-9317-8_5

[27] V. B. Savant and R. D. Kasar, "A review on network security and cryptography," *Research Journal of Engineering and Technology*, vol. 12, no. 4, pp. 110–114, 2021. [Online]. Available: https://doi.org/10.12691/iteces-3-1-1

[28] S. Bourougaa-Tria, F. Mokhati, H. Tria, and O. Bouziane, "Spubbin: Smart public bin based on deep learning waste classification an iot system for smart environment in algeria," *Informatica*, vol. 46, no. 8, 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i8.4331

[29] M. A. Al-Shareeda, S. Manickam, B. A. Mohammed, Z. G. Al-Mekhlafi, A. Qtaish, A. J. Alzahrani, G. Alshammari, A. A. Sallam, and K. Almekhlafi, "Cm-cppa: Chaotic map-based conditional privacy-preserving authentication scheme in 5g-enabled vehicular networks," *Sensors*, vol. 22, no. 13, p. 5026, 2022. [Online]. Available: https://doi.org/10.3390/s22135026

[30] H. Kaur and A. Kaur, "An empirical study of aging related bug prediction using cross project in cloud oriented software," *Informatica*, vol. 46, no. 8, 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i8.4197

[31] K. McKay, L. Bassham, M. Sönmez Turan, and N. Mouha, "Report on lightweight cryptography," National Institute of Standards and Technology, Tech. Rep., 2016. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8114.pdf

[32] M. A. Al-Shareeda, S. Manickam, B. A. Mohammed, Z. G. Al-Mekhlafi, A. Qtaish, A. J. Alzahrani, G. Alshammari, A. A. Sallam, and K. Almekhlafi, "Chebyshev polynomial-based scheme for resisting side-channel attacks in 5g-enabled vehicular networks," *Applied Sciences*, vol. 12, no. 12, p. 5939, 2022. [Online]. Available: https://doi.org/10.3390/app12125939

[33] S. Rajesh, V. Paul, V. G. Menon, and M. R. Khosravi, "A secure and efficient lightweight symmetric encryption scheme for transfer of text files between embedded iot devices," *Symmetry*, vol. 11, no. 2, p. 293, 2019. [Online]. Available: https://doi.org/10.3390/sym11020293

[34] M. A. Al-Shareeda, S. Manickam, B. A. Mohammed, Z. G. Al-Mekhlafi, A. Qtaish, A. J. Alzahrani, G. Alshammari, A. A. Sallam, and K. Almekhlafi, "Cm-cppa: Chaotic map-based conditional privacy-preserving authentication scheme in 5g-enabled vehicular networks," *Sensors*, vol. 22, no. 13, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/13/5026

[35] H. Ran, "Methodology for interval-valued intuitionistic fuzzy multiple attribute decision making and applications to performance evaluation of sustainable microfinance groups lending," *Informatica*, vol. 46, no. 8, 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i8.4355

[36] S. Nie, "Evaluation of innovative design of clothing image elements using image processing," *Informatica*, vol. 46, no. 8, 2022. [Online]. Available: https://doi.org/10.31449/inf.v46i8.4250

[37] M. Matsui, "Linear cryptanalysis method for des cipher," in *Workshop on the Theory and Application of of Cryptographic Techniques*. Springer, 1993, pp. 386–397. [Online]. Available: https://doi.org/10.1007/3-540-48285-7_33

[38] M. Shoeb and V. K. Gupta, "A crypt analysis of the tiny encryption algorithm in key generation," *International Journal of communication and computer Technologies*, vol. 1, no. 1, pp. 9–9, 2019. [Online]. Available: https://www.bibliomed.org/?mno=302643835

[39] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers, "Simon and speck: Block ciphers for the internet of things," *Cryptology ePrint Archive*, 2015. [Online]. Available: https://csrc.nist.gov/csrc/media/events/lightweight-cryptography-workshop-2015/documents/papers/session1-shors-paper.pdf

[40] B. Singh, G. Athithan, and R. Pillai, "On extensions of the one-time-pad," *Cryptology ePrint Archive*, 2021. [Online]. Available: https://eprint.iacr.org/2021/298.pdf

Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection

Sadri Alija [1], Edmond Beqiri [2*], Alaa Sahl Gaafar [3], Alaa Khalaf Hamoud [4]
[1] Faculty of Business and Economics, South East European University, North Macedonia.
[2] University of Peja "Haxhi Zeka" – Peja, Kosovo.
[3] Department of Educational Planning, Directorate of Education in Basrah, Iraq.
[4] Department of Computer Information Systems, University of Basrah, Iraq.
Email: s.aliji@seeu.edu.mk, edmond.beqiri@unhz.eu, alaasy.2040@gmail.com, alaa.hamoud@uobasrah.edu.iq.

Keywords: supervised machine learning, feature selection, wrapper, particle swarm optimization, info gain, SMOTE

Received: November 14, 2022

For learning environments like schools and colleges, predicting the performance of students is one of the most crucial topics since it aids in the creation of practical systems that, among other things, promote academic performance and prevent dropout. The decision-makers and stakeholders in educational institutions always seek tools that help in predicting the number of failed courses for the students. These tools can help in finding and investigating the factors that led to this failure. In this paper, many supervised machine learning algorithms will investigate finding and exploring the optimal algorithm for predicting the number of failed courses of students. An imbalanced dataset will be handled with Synthetic Minority Oversampling TEchinque (SMOTE) to get an equal representation of the final class. Two feature selection approaches will be implemented to find the best approach that produces a highly accurate prediction. Wrapper with Particle Swarm Optimization (SPO) will be applied to find the optimal subset of features, and Info Gain with ranker to get the most correlated individual features to the final class. Many supervised algorithms will be implemented such as (Naïve Bayes, Random Forest, Random Tree, C4.5, LMT, Logistic, and Sequential Minimal Optimization algorithm (SMO)). The findings show that the wrapper filter with SPO-based SMOTE outperforms the Info-Gain filter with SMOTE and improves the performance of the algorithms. Random Forest outperforms the other supervised machine learning algorithms with (85.6%) in TP average rate and Recall, and (96.7%) in ROC curve.

Povzetek: Opisana je metoda za napovedovanje uspeha študentov s pomočjo strojnega učenja.

1 Introduction

High-quality universities always require a great record of their students and the students are the main resource for them. The main concern for the universities is the performance of the students which is the base stone for building the top rate graduates and post-graduate students who will be the leaders of the nations and take responsibility of the economic and social growth of the society. Moreover, the main concerns for market employers are the performance of universities and students' academic performance due to its direct effect on the employment process and then employee productivity. So, the employers' demands are met by the graduated students who exert efforts in their academic journey. Student performance is measured by the learning assessment and the curriculum according to Usamah et al [1].

It is frequently important to be able to predict the behavior of future students to enhance the design of the curriculum and prepare the interventions for academic guidance and support. Machine learning (ML) is useful in this situation. ML approaches examine datasets, extract information, and then organize that information for eventual use. The primary goals of ML are to identify and extract patterns from recorded data by using a variety of techniques and algorithms [2]. Numerous algorithms exist and are used with educational data, including supervised algorithms such as Decision Tree (DT) and Naive Bayes (NB), and unsupervised algorithms such as K-Nearest Neighbor (KNN), and Neural Network (NN). Such algorithms forecast patterns, upcoming trends, and behaviors, enabling businesses to make informed, proactive decisions mining. This paper's major goal is to predict student performance using Supervised ML based on an imbalanced dataset and wrapper feature selection. The following section sheds light on related previous studies, then followed by the methodology and the concluded points, and future work.

2 Literature review

High quality universities always require the great record of their students where the students are the main resource for them. The main concern for the universities is the performance of the students which is the base stone for building the top rate graduates and post-graduates students who will be the leaders of the nations and take the responsibilities of the economic and social growth of the

The concept of data mining techniques can be implemented and applied in the educational field to improve our comprehension of the learning process, with a particular emphasis on the identification, extraction, and evaluation of factors linked to students' learning processes [3]. ML algorithms enable users to categorize and summarize associations discovered throughout the mining process as well as examine data from different perspectives. Bhardwaj and Pal in [4] explore the performance of the students by taking a sample of 300 undergraduate students' row records from the department of computer application from different institutions in Dr. R. M. L. Awadh University, India. The Bayesian classifiers are utilized on 17 features where the researchers found that there is a strong correlation between student action and other factors such as (living location, the academic background of the mother, senior secondary exam, the status, and the annual outcome of the student's family).

Next, in the same university, Pandey and Pal [5] selected 600 students to implement the model based on Bayes classifier to classify the background qualification, category, and language. While Hijazi and Naqvi in [6] have selected 300 students (75 female, and 225 male) from different colleges in Pakistan's Punjab University to explore and investigate student performance. Based on the linear regression, they found that there are many factors that affected the student's performance such as the attitude toward the class they attend, the time spent in studying after college, the mothers' ages, the income of their families, and the educational level of their mothers (where the performance is strongly affected by it). Khan in [7], explored the performance by building a model based on a clustering approach using 400 rows of student data from Aligarh Muslim University's senior secondary school in Aligarh, India. The main goal of the study is to determine the predictive value of different measures such as personality, cognition, and demographic variables that affect success at a higher level of secondary school. The outcomes of the study found that females with socioeconomic status scored higher performance, whereas males with low socioeconomics had higher performance in the science stream.

In the next case study [8], Kovacic implemented a data mining model for determining the educational enrollment data in New Zealand to predict the performance of the students. Chi-square automatic interaction detection (CHAID) and Classification and Regression (CART) algorithms are utilized to categorize the successful and failed students. The algorithms did not produce promising accuracies where they predicted the results with (59.4, and 60.5 respectively). The other case study is implemented by Galit [9] where the learning behavior is examined to predict the students' outcomes and alert the students to the critical status before the final exam. The final study [10] is proposed by Al-Radaideh, where the model is implemented to predict the students' final grades in C++ course for the students enrolled in the Yarmouk university in 2005, in Jordan. NB, DT (ID3, and C4.5) are utilized to predict the grades where the DT has outperformed the NB in prediction.

In our proposed model, the problem of imbalanced dataset is handled and the effect of handing this problem is observed by implementing different machine learning algorithms (supervised and unsupervised). The effect of handling imbalanced dataset is also observed by implementing feature selection which has the direct effect on the result accuracies.

3 Methodology

The model implementation framework is depicted in Figure 1, which consists of five steps starting with data preprocessing and ending with the model evaluation. The step of attribute feature selection (FS) is implemented by a single FS and a subset FS to find the effect of each step on the result accuracies. SMOTE filter is applied then, where it is followed by implementing supervised ML algorithms.

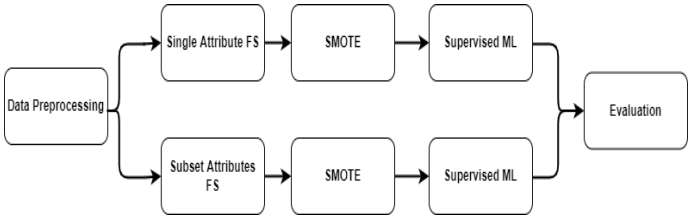


Figure 1: Model framework

3.1 Dataset reliability

A questionnaire is adopted in this study to build the model where Google Forms is used to build the questionnaire and collect undergraduate students' answers from both of Faculty of Contemporary Sciences and Technologies (CST) and the Faculty of Business and Economics (FBE) in South East European University (SEEU) in North Macedonia (RNM). The aim of this study is to find the optimal DT in predicting student performance based on the conceptual framework that was implemented by researchers in [11]. The aim of the framework is to find the hidden patterns that may affect and correlate with the performance of the students and provide suggestions to enhance and improve the performance. Many questions related to many factors are found in the questionnaire, such as academic behavior, health, finance, time planning, self-development, social relationships, and achieving goals. The questionnaire in [11] lists the factors and the questions related to each question, where the answer for most of the questions was on a 5-point Likert scale (from

1 to 5) which represented the formal answers (from "Strongly Disagree" to "Strongly Agree").
The dataset of the questionnaire involves 141 rows of respondents. The dataset reliability is required to measure the overall consistency of the dataset. The measure of reliability which describes consistency can be confirmed to have a high level if it produces similar results under consistent conditions. The most frequent measure in statistics is the coefficient alpha, which is used to calculate the internal consistency of the independent variables of the study. The coefficient's alpha for the dataset is 0.93. This value indicates an excellent internal consistency of the dataset reliability [12][13]. The applied tool for this model is Weka 3.8.5 and the system specifications are (RAM 8GB, HARD 35.5GB free, OS Win7 Pro).

Table 1: Dataset reliability

Number of Respondents	Number of Features	Coefficient's Alpha	% of Respondents
141	58	0.93	100%

3.2 Feature selection (FS)

FS approach can be considered as a form of data reduction where features are reduced and only the correlated features remain. The main goal of FS methods is finding the optimal subset of features or the highly correlated features that have a direct effect or may affect the final class(s). Due to the number of attributes in our dataset (57), it is required to find the most correlated attributes or features that can be utilized in the next steps to get more accurate results in classification [14]. Two approaches are applied in our model (Wrapper with Particle Swarm Optimization (PSO)) and (Info-Gain Attribute Evaluator).

- **Wrapper method**

The Wrapper method evaluates the subset of attributes according to the classifier performance for both supervised algorithms (such as DT, SVM, and NB) and unsupervised algorithms (such as clustering). For each subset, the evaluation process is repeated while the search strategy determines the subset generation. The wrapper method is slower than the filter in finding good subsets because it depends on resource demands for the algorithm of modeling. Due to using real modeling algorithms, the wrapper method is proven empirically to produce better feature subsets [15].

- **Particle swarm optimization (PSO)**

Kennedy and Eberhart in 1995 proposed one of the evolutionary computation techniques based on social behavior such as fish schooling and bird flocking. The basic idea behind PSO underlines that the population-social interaction optimizes knowledge where the thinking is personal and social. The solutions are represented by particles, while particles are represented by vectors that have positions in the search space. Each vector xi=(xi1,xi2,…xiD) Where D is the search space dimensionality. To search for the optimal solutions, the particles move in the search space. According to that, each particle has a velocity that can be represented by vi where vi takes the values (vi1,vi2,….,viD). The particle updates its location and velocity during the movement, and this update is performed according to the neighbors and their own experience. Two values of positions are recorded, the best which represents the best previous personal position of the particle, and gbest is the best-obtained position by the population. The following equation is used to update the position and velocity:

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (1)$$

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1 * (p_{id} - x_{id}^{t}) + c_2 * r_2 * (p_{gd} - x_{id}^{t}) \quad (2)$$

Where t is the tth iteration in the evolutionary process while d represents the d dimension in the search space where d belongs to D. The weight w it controls the previous velocity impact on the current velocity impact. The acceleration constants c_1, c_2 are random values in the range (0 to 1), p_{id} and p_{gd} represent the elements of *pbest*, *gbest* alternatively in the dimension dth. *vmax* is the maximum velocity where $v_{id}^{t+1} \in$ [−*vmax*, *vmax*]. The algorithm will stop when the predefined criterion of fitness is met with a good fitness value or a predefined number of maximum iterations [16][17].

- **Info gain**

In this feature selection evaluator, the information of each class is estimated to evaluate the attribute. The method used in this evaluator is minimum description-length-based discretization where the attributes are binarized or discretized. In this method, the missing values are either regarded as separate values or distributed the values among other values according to the frequencies. As the value of the feature is absent, the decrease in entropy is measured. For the multiclass attribute, the InfoGain evaluator has reported the best performance. The generalized form of the nominal values is taken from the nominal attribute. Info Gain is measured by the decrease of X entropy that is caused by Y which is represented by:

$$IG(X|Y) = H(X) - H(X|Y) \quad (3)$$

According to this measurement, (Y) feature can be considered as more correlated to (X) feature if (IG(X | Y) > IG(Z | Y). IG normalized the values that fall within the range (0 to1), where (1) value indicates that the predicted value is completely correct and (0) value indicates that (X) feature is independent of (Y) feature. For the nominal and continuous features, the Entropy can be applied in order to determine the correlation between continuous and nominal features [18][19][20][21].
The Wrapper filter with SPO is applied to find and explore the most correlated subsets of features that make the highly accurate results for each supervised algorithm. Wrapper as a subset of attributes evaluator is applied for each supervised classifier individually. In this step, different subsets of features are found for each classifier where the SPO is selected as a search method to improve the speed of search for features subsets. In order to find

the effect of wrapper evaluator, Info Gain evaluator is applied to find the features with high correlations with the final class and to find how wrapper and Info Gain affect the result algorithms accuracies of the algorithms. Table 2 shows the most correlated features (subset) after applying wrapper with SPO for each algorithm and Info Gain with Ranker.

Table 2: Selection of attributes

Feature Evaluator	Attributes
Wrapper (Random Forest) with SPO	1,5,6,7,8,9,10,12,13,14,16,17,18,27,33,36,44,49,52,53,56
Wrapper (NB) with SPO	5,8,14,18,25,31,42,48
Wrapper (Logistic) with SPO	2,4,5,6,11,13,17,31,35,48,51,52,53,54,57
Wrapper (SimpleLogistic) with SPO	1,4,5,6,8,9,11,15,17,23,26,27,28,31,32,34,42,44,46,50,52,53,55
Wrapper (SMO) with SPO	4,5,14,15,17,24,31,32,35,42,45,47,54,55,56,57
Wrapper (LMT) with SPO	1,2,4,5,6,7,8,9,11,14,15,17,19,20,21,23,25,26,27,28,32,34,41,42,44,49,52,53,55
Wrapper (J48) with SPO	5,7,13,22,23,24,26,31,35,42,45,46,52
Wrapper (Random Tree) with SPO	5,15,27,33,35,43,44,45,46,48,49
Info Gain with Ranker	5,57,19,18,21,17,20,22,15,23,26,25,24,16,14,28,7,4,3,2,6

3.3 Synthetic minority over-sampling technique (SMOTE)

The dataset is said to be imbalanced if the classes in the final class are not equally represented [22]. If the final class has the classes (1,2, and 3) and the representations of the classes are (10% for 1, 15% for 2 and 75% for 3) then the dataset is imbalanced. The imbalanced datasets are found in almost all sectors starting from the medical sector [23], telecommunications management [24], fraudulent telephone calls [25], and text classification [26]. The SMOTE approach creates "synthetic" examples, to oversample the minority classes or by replacing the samples. This approach has been inspired and proven its success by the recognition process of handwritten characters [27]. The generation of synthetic examples is performed based on the operating in the feature space rather than the data space. The data space will face certain operations to generate the training data. The process of oversampling is performed by taking each minority class attribute of the final class attribute and introducing new examples (synthetic) along the line segments which join all k classes if they are nearest neighbors. The selection of the k nearest neighbors is performed randomly according to the oversampling amount required. The synthetic samples generation is implemented by taking the difference between each sample with its neighbors, then the result difference is multiplied by a random number between 0 and 1; then the result obtained is added to the feature vector. This process effectively forces to make the minority class more generally, see Figure 2 [28].

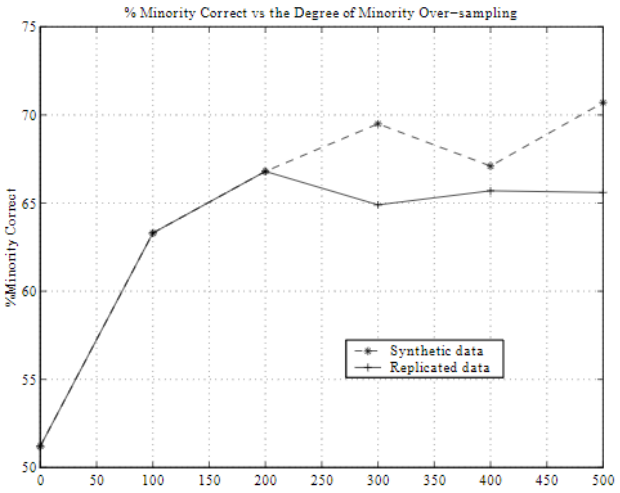


Figure 2: Comparison of number of minority correct for replicated oversampling and SMOTE for a dataset [28].

In our imbalanced dataset, the percentage of classes' representation is shown in Table 3. Class (3) takes only (4.3%) of the overall dataset, followed by classes (1, and 2) respectively with (21.3%, and 21.9%). The SMOTE filter in our model will be implemented on the classes (1,2, and 3) to make the dataset balanced and to get reliable performances of the algorithms. The SMOTE filter is applied to get equal representations of all classes.

Table 3: Classes representation

Class	Number of Rows	% of Representation
0	74	52.5%
1	30	21.3%
2	31	21.9%
3	6	4.3%

3.4 Supervised machine learning (ML)

In the proposed model, many supervised ML algorithms have been implemented to find the accurate algorithm for predicting the number of failed courses for the students. The algorithms fall in approaches such as (decision tree (DT) (Random Forest, Random Tree, LMT, and J48), naïve Bayes (Naïve Bayes, and Bayes Net), Logistic (Logistic and Simple Logistic), and Support Vector Machine (SMO)). DT is one of the supervised ML approaches that aim to build a training model to be used in predicting the final class attribute [29]. DT classifiers are widely used in different sectors and have proved their accuracies in the fields of education [11], [30][31], healthcare [32], wireless sensor networks [33], image processing [34][35], and disaster management [36][37]. There are many types of algorithms and the most used algorithms are (Random Forest, CART, Iterative Dichotomies 3 (ID3), and Successor of ID3 (C4.5 or J48) [38][39]. DT is used in the field of classification (predicting the categorical values) and regression (predicting the continuous values) [40]. Random Forest (which was proposed by L. Breiman in 2001) is a general-purpose regression and classification approach that works on the principle of aggregating the predictions by calculating the predictions averages and shows excellent

performance when the variables numbers is larger than the number of the observations [41]. In logistic model trees (LMT), logistic regression is utilized to select the attributes in a natural way by using stage-wise fitting. The logistic model in this approach is built on leaves by refining the leaves incrementally at the higher level of the tree [42].

SVM is an ML algorithm that falls under the supervised learning algorithm [43], as it is one of the data-based algorithms used to solve classification problems. It is considered one of the most important algorithms to accomplish that task (solving classification problems) [44]. Support Vector Machine has a vector support processing approach in which many questions are answered depending on the understanding and knowledge of the problem and how to design it. Moving to the real world, we find that the Support Vector Machine algorithm was used to find solutions to many problems in this world, including face recognition, detection, hand lines, and others [45]. In order to understand the SVM algorithm, it is necessary to understand its main terminology, the maximum-margin hyperplane, the separating hyperplane, the soft margin, and the kernel function [43]. SVM can be classified into two types: Linear SVM, and Non-Linear SVM. Linear SVM is an algorithm used when the data can be separated into two groups in a linear way by using a straight line where the data can be called as linearly separable, in addition to that the classifier is described as SVM classifier. Non-Linear SVM is an algorithm used when the data cannot be separated in a linear manner, and thus a straight line cannot be used to separate the data into two categories. To compensate for this, another thing called the kernel trick is used, through which we define the data in a higher dimension to be separated using some mathematical functions.

Regression is considered a simple type of ML algorithm. It is considered a supervised learning algorithm. These algorithms are used in a wide range to find a relationship between the continuous predictor and response variables. It is considered a way to measure the relationships between response variables and continuous predictors [46]. An example of this is the linear regression algorithm, which is one of the supervised learning algorithms, where this algorithm simulates the mathematical relationship between variables. It attempts to find relationships between independent variables (input data) and dependent variables (result, and forecast). It works to find continuous or numerical variables by predicting that as it assumes that the relationships between the predicted variables and the goal to be reached are linear, such as sales, age, and product price. The regression may be linear or curvilinear, so it must pass through all data points to reach the target prediction so that if the measurement is made between the data points and the regression line, the result is minimal.

In order to solve classification problems, a logistic regression algorithm was built, which is one of the supervised learning algorithms, where the results are always binary, not devoid of one of the two values, either 1 or 0, success or failure, rain or no rain; its working principle is probability. Logistic regression is used in the analysis of binary outcomes, or as it is said that they are two-level, or whose levels are opposite [47]. A characteristic of logistic regression is that its predictions are deterministic and have the ability to adapt to multiple predictions. This is necessary for the analysis of observational data when adjustment is useful to avoid differences in the totals to be compared [48]. Logistic regression is used to reach the highest weighting of a variable in the event that there is more than one variable. Thus, it is similar in terms of multiple linear regression and is inconsistent with it that the response variable has only a binomial, and as a result, each variable is considered to have an impact on the likelihood ratio of any expected event. Hence, it has the advantage that it can avoid confusing influences by analyzing the correlations of all variables at the same time [49].

NB is considered one of the supervisor learning algorithms; it is based on Bayes' rule together with additional to strong assumptions attributes that are categorically and conditionally independent [50]. Then it is used for solving classification problems. This algorithm assumes conditional independence of traits; so it is rarely true in the real world, which has made the competitive performance of this algorithm a lot of attention and surprising [51]. The Naïve Bayes algorithm is used in a wide range of applications, including article classification and spam filtering. Naïve Bayes Classifier is able to build ML model through which we get fast predictions. The hypothesis states that the independence between every two features, so the naïve Bayes classifier calculates the probability of belonging to a certain class. As a product of simple probabilities resulting from assumed Naïve independence. The hypothesis states that there is independence between each of the two features, so the Naïve Bayes classifier computes the probability of a particular instance belonging to a particular class. If we assume that the described is described by a vector x of attributes and the target of the class is the element y, then we can express the conditional probability $p(y|x)$ as the product of the simple probabilities resulting from the assumed naïve Bayes independence [52].

Bayesian networks are considered probabilistic models that depend mainly on non-periodic direct graphs. These models are causal relationships between their variables, and their structure represents the combination of previous knowledge and target data. They are also called belief networks as they belong to probabilistic graphical models, and knowledge can be represented in uncertain domains through the use of their graphical structures. It is observed by looking at its graphs, where nodes represent random variables, while arrows between nodes (variables) represent probabilistic dependencies. In most cases, generally accepted statistical methods are used to estimate these conditional dependencies. Hence, we can say that Bayesian networks combine graphs and statistics as well as computer science and probability theory [53]. Also, Bayesian networks are used to perform causal logic and predict risks. In addition, there are many advantages if we compare it with the methods used in regression methods [54]. One of Bayes Network's products is the modeling language in addition to the inference algorithms associated with random domains. Experiments have proven a lot of

success when used in medium-sized applications. But if Bayesian networks are used in areas that are relatively complex or large domains, then these networks will use the task of modeling, which is somewhat similar to programming using logic circuits [55].

3.5 Model evaluation

The evaluation process of algorithms is performed based on the confusion matrix, see Figure 3. The class value of True Negative (TN) is the predicted class as (NO) and it is (NO), while the class value of False Positive (FP) is the class when it is predicted as (YES) and it is (NO). False Negative (FN) class value is the class when it is predicted as (NO) and it is (YES) while True Positive (TP) class value is the class when it is predicted as (YES) and it is (YES).

	Predicted (YES)	Predicted (NO)
Actual (YES)	True Positive (TP)	False Negative (FN)
Actual (NO)	False Positive (FP)	True Negative (TN)

Figure 3: Confusion matrix.

Based on the above matrix, the performance criteria are:

$$Sensevity\ or\ TP = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity\ or\ FP = \frac{FP}{FP+TN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

The sensitivity or recall is a measurement of the truly predicted cases and measures the relevance of TP with FN. The more the TP rate, the more accurate the predicted cases and the more accurate the classification algorithm. The specificity or FP rate is the false alarm rate that measures the incorrectly predicted cases. The more FP, the more predicted incorrect cases. The precision represents the relevant cases among the predicted cases [29]–[31].

Table 4: Algorithms performance after wrapper with SPO.

Algorithm	TP Rate	FP Rate	Precision	Recall
LMT	0.766	0.078	0.762	0.766
Random Forest	**0.856**	**0.049**	**0.857**	**0.856**
Random Tree	0.697	0.100	0.695	0.697
NB	0.717	0.094	0.711	0.717
Logistic	0.727	0.091	0.729	0.727
Simple Logistic	0.773	0.075	0.770	0.773
SMO	0.757	0.081	0.752	0.757
J48	0.796	0.067	0.796	0.796

Table 4 lists the performance evaluation of supervised algorithms after implementing Wrapper with SPO. The algorithms are implemented after removing the uncorrelated features where the Wrapper base classifier is the supervised algorithm. Then, the SMOTE filter is applied to get equal representations of classes for the final class. RF algorithm outperforms the other supervised algorithms with (85.6% in TP rate and Recall), (4.9% in FP rate) and (85.7%) in precision. C4.5 (J48) algorithm comes in the second rank with (79.6% in TP rate and Recall), (6.7% in FP rate), and (79.6%) in Precision. NB comes in the last rank with (71.7% in TP rate and Recall), (9.4%) in FP rate, and (71.1%) in Precision.

Table 5: Algorithms performance after info gain evaluator.

Algorithm	TP Rate	FP Rate	Precision	Recall
LMT	0.750	0.083	0.749	0.750
Random Forest	**0.836**	**0.054**	**0.835**	**0.836**
Random Tree	0.701	0.099	0.696	0.701
NB	0.678	0.107	0.678	0.678
Logistic	0.737	0.087	0.735	0.737
Simple Logistic	0.707	0.097	0.701	0.707
SMO	0.734	0.088	0.730	0.734
J48	0.753	0.082	0.750	0.753
Bayes Net	0.750	0.083	0.753	0.750

Table 5 depicts the performance criteria of supervised ML algorithms after implementing Info Gain. The algorithms are implemented after removing the uncorrelated features (36 features), then the SMOTE filter is applied to get equal representations of classes for the final class. RF algorithm outperforms the other supervised algorithms with (83.6% in TP rate and Recall), (5.4% in FP rate) and (83.5%) in precision. C4.5 (J48) algorithm comes in the second rank with (75.3% in TP rate and Recall), (8.2% in FP rate), and (75%) in Precision. NB comes in the last rank with (67.8% in TP rate and Recall), (10.7%) in FP rate, and (67.8%) in Precision.

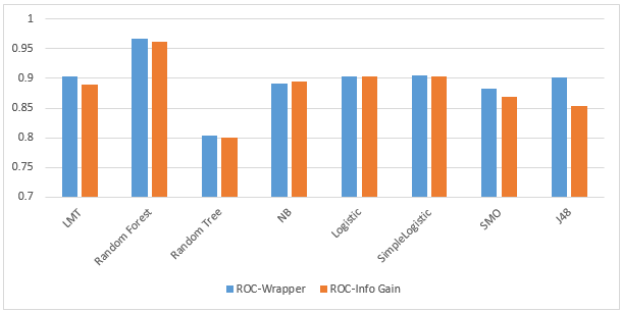


Figure 4: ROC of algorithms with wrapper and info gain.

One of the performance criteria that determines the optimal classifiers is the Receiver Operating Characteristic (ROC) curve, where ROC is considered one of the standard techniques that summarize classifier performance over a range of tradeoffs between TP and FP error rates [32][28]. As much as the ROC is closer to 1, as much as the classifier is accurate. Based on Figure 4, the RF classifier is the optimal classifier among all other classifiers with (96.7%) ROC when the wrapper with SPO

is implemented. The ROC is (96.1%) for the same classifier when Info Gain is implemented. The figure shows that ROCs for all algorithms are enhanced after implementing a wrapper evaluator with SPO. NB is the only classifier that has (89.1%) ROC when implementing wrapper and (89.5%) with Info Gain Evaluator.

4 Conclusions and future works

The imbalanced dataset faced many techniques and approaches to solve the minority and majority class problems related to the final class. In our model, the imbalanced dataset has multi-values in the final class which is required to handle this problem using SMOTE filter. In our model, the step of feature selection is performed two ways, the first one is by applying wrapper evaluator with SPO as a search method to find subsets of attributes that may affect and be correlated with the final class, and the second one by applying Info Gain as an evaluator with ranker as a search method to find the features with most correlation with the final class. After finding the most correlated features or feature subsets using evaluators, the uncorrelated features are removed and the SMOTE filter is applied to produce a balanced dataset and to make the multi-values classes equally represented. Many supervised ML algorithms are applied such as (NB, RF, Random Tree, LMT, J48, Logistic, Simple Logistic, and SMO). The performance evaluation of the algorithms shows that using the wrapper with the classifiers and SPO as a search method outperforms the Info-Gain evaluator. RF algorithm outperforms other algorithms in predicting students' performance and the number of failed courses. The model can be updated by predicting the students' status whether will fail or pass the final class. The features will be explored and investigated using different filters and classifiers to find the features with the most correlations with students' failure.

References

[1] U. Bin Mat, N. Buniyamin, P. M. Arsad, and R. A. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," in 2013 IEEE 5th International Conference on Engineering Education: Aligning Engineering Education with Industrial Needs for Nation Development, ICEED 2013, 126-130, 2014. https://doi.org/10.1109/iceed.2013.6908316.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., vol. 17, no. 3, 37-37, 1996.

[3] A. El-Halees, "Mining Students Data To Analyze Learning Behavior : a Case Study Educational Systems," Work, 2008.

[4] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for performance improvement using classification," World J. Comput. Appl. Technol., vol. 2, no. 2, 2014. https://doi.org/10.13189/wjcat.2014.020203

[5] U. K. Pandey and S. Pal, "Data Mining: A prediction of performer or underperformer using classification," arXiv Prepr. arXiv1104.4163, 2011.

[6] S. M. M. Syed Tahir Hijazi & Raza Naqvi, "Factors affecting students' performance: A case of private colleges," Bangladesh e-Journal Sociol., vol. 3, no. 1, pp. 1–10, 2006.

[7] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream," J. Soc. Sci., vol. 1, no. 2, 2005. https://doi.org/10.3844/jssp.2005.84.87

[8] Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," in Proceedings of the 2010 InSITE Conference, 2010. https://doi.org/10.28945/1281

[9] G. (Univ T. A. Ben-Zadok, R. (Univ T. A. Mintz, A. (Univ T. A. Hershkovitz, and R. (Univ T. A. Nachmias, "Examining online learning processes based on log files analysis: A case study," Res. Reflections Innov. Integr. ICT Educ. Proc. Fifth Intertnational Conf. Multimdeia ICT Educ., no. 2, 2009.

[10] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[11] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," Int. J. Interact. Multimed. Artif. Intell., 2018. https://doi.org/10.9781/ijimai.2018.02.004

[12] B. Carson, "The transformative power of action learning," Chief Learn. Off. Retrieved, 2017.

[13] U. Sekaran and R. Bougie, Research methods for business: A skill building approach. john wiley & sons, 2016.

[14] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," Computers in Biology and Medicine, vol. 112. 2019. https://doi.org/10.1016/j.compbiomed.2019.103375

[15] Y. Kim, W. N. Street, and F. Menczer, "Evolutionary model selection in unsupervised learning," Intell. Data Anal., vol. 6, no. 6, 2002. https://doi.org/10.3233/ida-2002-6605

[16] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," IEEE Trans. Cybern., vol. 43, no. 6, 2013. https://doi.org/10.1109/tsmcb.2012.2227469

[17] Y. Shi and R. Eberhart, "Modified particle swarm optimizer," in Proceedings of the IEEE Conference on Evolutionary Computation, ICEC, 1998. https://doi.org/10.1109/icec.1998.699146

[18] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in Proceedings, Twentieth International Conference on Machine Learning, 2003, vol. 2.

[19] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, Fourth Ed., 2016. https://doi.org/10.1016/b978-0-12-374856-0.00010-9

[20] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993.

[21] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," Data Min. Knowl. Discov., vol. 6, no. 4, 2002.

[22] F. Provost and T. Fawcett, "Robust classification for imprecise environments," Mach. Learn., vol. 42, no. 3, 2001.

[23] A. S. Desuky, A. H. Omar, and N. M. Mostafa, "Boosting with crossover for improving imbalanced medical datasets classification," Bull. Electr. Eng. Informatics, vol. 10, no. 5, 2021. https://doi.org/10.11591/eei.v10i5.3121

[24] J. Xiao, L. Xie, C. He, and X. Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," Expert Syst. Appl., vol. 39, no. 3, 2012. https://doi.org/10.1016/j.eswa.2011.09.059

[25] C. Lu, S. Lin, X. Liu, and H. Shi, "Telecom fraud identification based on ADASYN and random forest," in 2020 5th International Conference on Computer and Communication Systems, ICCCS 2020, 2020. https://doi.org/10.1109/icccs49078.2020.9118521

[26] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," in Procedia Computer Science, 2019, vol. 159. https://doi.org/10.1016/j.procs.2019.09.229

[27] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation method," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 5, 1997. https://doi.org/10.1109/34.589216

[28] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 3 https://doi.org/10.1613/jair.953 21–357, 2002.

[29] M. A. Kumar and A. J. Laxmi, "Machine Learning Based Intentional Islanding Algorithm for DERs in Disaster Management," IEEE Access, vol. 9, 2021. https://doi.org/10.1109/access.2021.3087914

[30] A. K. Hamoud, "Selection of Best Decision Tree Algorithm for Prediction and Classification of Students' Action," Am. Int. J. Res. Sci. Technol. Eng. Math., vol. 16, no. 1, pp. 26–32, 2016.

[31] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in IOP Conference Series: Materials Science and Engineering, 2020, vol. 928, no. 3, p. 32019. https://doi.org/10.1088/1757-899x/928/3/032019

[32] T. Saba, I. Abunadi, M. N. Shahzad, and A. R. Khan, "Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types," Microsc. Res. Tech., vol. 84, no. 7, 2021. https://doi.org/10.1002/jemt.23702

[33] I. A. Najm, A. K. Hamoud, J. Lloret, and I. Bosch, "Machine Learning Prediction Approach to Enhance Congestion Control in 5G IoT Environment," Electronics, vol. 8, no. 6, p. 607, May 2019. https://doi.org/10.3390/electronics8060607

[34] J. Chen, Y. Lian, and Y. Li, "Real-time grain impurity sensing for rice combine harvesters using image processing and decision-tree algorithm," Comput. Electron. Agric., vol. 175, 2020. https://doi.org/10.1016/j.compag.2020.105591

[35] I. S. Masad, A. Al-Fahoum, and I. Abu-Qasmieh, "Automated measurements of lumbar lordosis in T2-MR images using decision tree classifier and morphological image processing," Eng. Sci. Technol. an Int. J., vol. 22, no. 4, 2019. https://doi.org/10.1016/j.jestch.2019.03.002

[36] S. Khatoon et al., "Development of social media analytics system for emergency event detection and crisismanagement," Comput. Mater. Contin., vol. 68, no. 3, 2021. https://doi.org/10.32604/cmc.2021.017371

[37] H. Li, D. Caragea, C. Caragea, and N. Herndon, "Disaster response aided by tweet classification with a domain adaptation approach," J. Contingencies Cris. Manag., vol. 26, no. 1, 2018. https://doi.org/10.1111/1468-5973.12194

[38] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," Shanghai Arch. Psychiatry, vol. 27, no. 2, 2015.

[39] N. Mahdi Abdulkareem and A. Mohsin Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," Int. J. Sci. Bus., vol. 5, no. 2, 2021.

[40] S. M. Rasoolimanesh, M. Wang, J. L. Roldán, and P. Kunasekaran, "Are we in right path for mediation analysis? Reviewing the literature and proposing robust guidelines," J. Hosp. Tour. Manag., vol. 48, 2021. https://doi.org/10.1016/j.jhtm.2021.07.013

[41] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, no. 2, 2016. https://doi.org/10.1007/s11749-016-0481-7

[42] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," Mach. Learn., vol. 59, no. 1, pp. 161–205, 2005. https://doi.org/10.1007/s10994-005-0466-3

[43] W. S. Noble, "What is a support vector machine?" Nature Biotechnology, vol. 24, no. 12. 2006. https://doi.org/10.1038/nbt1206-1565

[44] T. Joachims, "Svmlight: Support vector machine," SVM-Light Support Vector Mach. http//svmlight.joachims.org/, Univ. Dortmund, vol. 19, no. 4, 1999.

[45] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and

non-linear pattern classification," in Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019, 2019. https://doi.org/10.1109/iss1.2019.8908018

[46] K. Park, R. Rothfeder, S. Petheram, F. Buaku, R. Ewing, and W. H. Greene, "Linear regression," in Basic Quantitative Research Methods for Urban Planners, 2020. https://doi.org/10.4324/9780429325021-12

[47] A. J. Scott, D. W. Hosmer, and S. Lemeshow, "Applied Logistic Regression.," Biometrics, vol. 47, no. 4, 1991. https://doi.org/10.2307/2532419

[48] B. R. Kirkwood and J. A. C. Sterne, Essential Medical Statistics. 2003.

[49] S. Sperandei, "Understanding logistic regression analysis," Biochem. Medica, vol. 24, no. 1, 2014. https://doi.org/10.11613/bm.2014.003

[50] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes.," Encycl. Mach. Learn., vol. 15, pp. 713–714, 2010. https://doi.org/10.1007/978-0-387-30164-8_576

[51] H. Zhang, "The optimality of naive Bayes," Aa, vol. 1, no. 2, p. 3, 2004.

[52] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes," PLoS One, vol. 9, no. 1, p. e86703, 2014. https://doi.org/10.1371/journal.pone.0086703

[53] J. Pearl, "Bayesian networks," 2011.

[54] P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, and M. J. Druzdzel, "Bayesian networks for risk prediction using real-world data: a tool for precision medicine," Value Heal., vol. 22, no. 4, pp. 439–445, 2019. https://doi.org/10.1016/j.jval.2019.01.006

[55] D. Koller and A. Pfeffer, "Object-oriented Bayesian networks," arXiv Prepr. arXiv1302.1554, 2013.

[56] A. Khalaf et al., "Supervised Learning Algorithms in Educational Data Mining: A Systematic Review," Southeast Eur. J. Soft Comput., vol. 10, no. 1, pp. 55–70, 2021.

https://doi.org/10.31449/inf.v47i1.4421 Informatica **47** (2023) 21–42 **21**

On Integrating Multiple Restriction Domains to Automatically Generate Test Cases of Model Transformations

Thi-Hanh Nguyen and Duc-Hanh Dang*
Department of Software Engineering,
VNU University of Engineering and Technology, Hanoi, Vietnam
E-mail: hanhit@hnue.edu.vn, hanhdd@vnu.edu.vn
*Corresponding author

Keywords: transformation testing, black-box testing, classifying term, model finding, OCL

Received: September 24, 2022

Testing model transformations poses several challenges, one of which is how to automatically generate effective test suites. A promising approach for this is to employ equivalence partitioning, a well-known technique for software testing. Specifically, in order to generate effective test suites, current works in literature often focus on exploiting either the structural aspects of models or transformation contracts for partition analysis. However, for the aim, they focus on only a single restriction source such as metamodels, contracts of the transformation, and domain-expert knowledge. To increase the effectiveness of generated test suites, partitioning techniques should be performed on a combination of various restriction sources. This paper introduces a method to generate test models on such a multi-domain of restrictions. The method also allows the tester to flexibly select and combine constraints to create a unified restriction for different strategies and objectives in model transformation testing. We developed a support tool based on the UML-based Specification Environment (USE) and performed experiments on several transformations to point out the effectiveness of our method.

Povzetek: Opisana je metoda preverjanja programske kode na osnovi multi-modalnih omejitev posameznih delov.

1 Introduction

Model transformations are the pillars of Model-Driven Engineering (MDE). Testing has been an effective technique to ensure the quality of model transformations which is the key to successfully realizing MDE in practice. This discipline consists of the following main tasks: synthesizing models as test data that are referred to as *test models*, performing the transformation, and verifying the output results. Until now, how to synthesize automatically and effectively *test models* for model transformations is still challenging.

The test model generation is the synthesis of models from different restriction sources including syntactic and semantic domains of source and target models. Such restriction domains often have complex structures and semantics that make it difficult to automate the generation. To the best of our knowledge, there are typical restriction domains in the context of MDE as follows. First, for a so-called *source metamodel coverage*, as explained in [1, 2, 3, 4, 5, 6], test models could be generated by applying the well-known testing technique *equivalence partitioning* that splits the input metamodel into equivalence partitions for selecting representative test models. Second, for a so-called *transformation specification coverage*, as proposed in [7, 8, 11], additional restrictions on source models could be derived from a transformation specification and taken as input contracts to generate test models. Within the works, input contracts of the transformation specification often are expressed as OCL conditions. Third, following the white-box testing approach, the works in [12, 13, 14, 15] focus on analyzing a model transformation implementation to build test suites using the notion of *transformation implementation coverage*. In addition, in interactive approaches, domain knowledge can support the test model selection. For example, based on the test objective, domain experts could choose representative values for the partition testing technique [1, 4, 16], or directly create examples for test models within test-driven development approaches, as explained in [18, 19].

Generating test models based on the analysis and synthesis of each single particular restriction domain can lead to a large duplication of test models, wasting testing time and effort. This highlights the need to generate test models from multiple restriction domains. However, realizing this need presents several challenges: (1) Constraints from multiple domains expressed in heterogeneous formalism need to be translated into a consistent and unified formalism to enable model synthesis. (2) The partition analysis technique is often employed to obtain representative test models since exhaustive testing is a non-trivial task, but defining a suitable partition on multiple restriction domains for different test strategies can be challenging. (3) The automatic generation of test models often requires manually defining (as input of

the solver) parameters for the testing environment as well as the other configuration information. This is challenging to automate this task.

This paper proposes a mechanism based on an integration of multiple restriction domains for a black-box testing approach to automatic generation of test models. Specifically, multi-domain restrictions that include (1) conditions for partitioning the metamodel and (2) transformation contracts are first translated into OCL conditions; and then taken as the input of a constraint solver for generating test models. For each common test strategy, a mechanism of combining OCL conditions should be established to define combinatorial partitions using logical operators. Moreover, a scope-value searching method needs to be incorporated to solve constraints and so that the set of generated test models has a reasonable size. The main contributions of this paper are summarized as follows:

– A method to automatically generate test models with multi-domain restrictions for effective model transformation testing.
– A mechanism to define suitable partitions for different test strategies.
– An OCL-based support tool and experimental results to show the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 surveys related works. Section 3 motivates this work with a transformation example. Section 4 outlines our approach. Section 5 explains restriction domains as a basis for a partition analysis and automatic generation of test models. Section 6 introduces several strategies to combine partitions in order to generate test models for different test objectives. Section 7 shows our tool support. Section 8 illustrates our testing method with several transformations and points out the effectiveness of our method. Section 9 explains threats to the validity of this work and discusses the results. This paper is closed with a conclusion and a discussion of future work.

2 Related work

In this section, we provide an overview of black-box testing approaches for model transformations and address the following research questions: (1) how to automatically generate test models using the partition analysis technique in a black-box testing approach, (2) how to construct test oracles that check test outputs to ensure quality properties; and (3) how to evaluate the quality of test suites in terms of one or more test objectives. First, a common basic idea of black-box testing approaches for transformations is to use metamodel and requirements specification as test basis, i.e., they are independent of transformation implementations. Within these approaches, the well-known testing technique *equivalence partition* [20] often is used to split the input data domain into equivalence partitions based on the test basis analysis and then to select a representative model for each partition. Fleurey et al. [1, 24] have proposed a partitioning technique based on the datatype of class attributes and the association end multiplicity within a UML class diagram representing the metamodel. Several other partitioning techniques for generating test models that conform to a metamodel as introduced in [2, 21, 4, 23, 5, 22, 6, 39, 9]. One of the main limitations of the metamodel-based partitioning approach is that the technique often generates a large number of test models, and generated test models tend to correspond to just only a subfragment of the source metamodel instead of the whole metamodel. To overcome this limitation, Fleurey introduces the notion of a so-called effective metamodel as the fragment of the source metamodel that is actually manipulated by the transformation. An effective metamodel can be defined by either examining the specification of transformations as explained in [1, 3] or statically analyzing their implementation as shown in [21].

The partitioning technique, as shown in [7, 10, 8, 9], can also be employed to specify requirements of model transformations. Following the research line, the works in [9, 39, 8] propose to derive partitioning conditions from a contract-based specification of the transformation. The specification of transformations within these approaches often includes preconditions and invariants as contracts on the input data domain of the transformation (i.e., corresponding to restrictions on source models). Partitioning conditions are then translated into either OCL constraints [7] or other first-order logic languages like Alloy [4, 23] for the automatic generation of test models using the model finding technique. For different test objectives, the works in [1, 6, 8, 4, 36] have proposed suitable techniques to improve the quality of test cases. Fleurey et al. [1] proposed the use of the Bacteriologic algorithm to optimize test suites. On analyzing metamodel-based partitioning, Fleurey et al. [1], Janabin et al. [6], Gogolla et al. [8], and Sen et al. [4] proposed using representative values provided by domain experts or testers. Similarly, Sen et al. [4] proposed combining different knowledge domains and uniformly representing them as constraints in Alloy. Cabot et al. [36] proposed a similar technique in which combinatorial partitioning conditions are represented in OCL.

Second, another major challenge for model transformation testing is how to predict desired expected outputs [1]. This research line can be divided into two groups: The first aims to predict the whole output model, i.e., making use of a *complete oracle function*, and the other aims to predict just part of desired target model, i.e., using a *partial oracle function*. The first approach (with complete oracle functions) would take the expected output model as a reference model and check if the actual output model conforms to this reference model, e.g., using model comparison as regarded in [32, 15, 17, 30, 34, 35]. For this aim, Addazi et al. [35] employs EMFCompare, whereas the works in [17, 15] design specific algorithms to compare models. Besides, Kolovos [33] employs the Epsilon Comparison

Table 1: Black-box approaches for test model generation (**MP**=> Metamodel-based Partitioning; **SP** => Specification-based Partitioning; **MF** => Model Finding; **Al** => Algorithm)

Reference	Test Model Generation			Restriction domains				Test adequacy	
	Partitioning	Representing patterns	Generating models	Metamodel	Specification	Transformation implementation	Domain Expert	Metamodel coverage	Specification coverage
Fleurey et al. [1]	MP	Pattern	AL	*				*	
Wang et al. [2, 21]	MP	Pattern		*		*		*	
Wu et al. [5, 22]	MP	OCL	AL	*			*	*	
Lamari [3]	MP, SP	Pattern		*		*		*	
Jahanbin et al. [6]	MP	Pattern	AL	*				*	*
Sen et al. [4, 23]	MP, SP	Alloy	MF	*	*		*		*
Guerra et al. [7]	SP	Pattern, OCL	MF	*	*				*
Burgueño et al. [9, 39]	MP	OCL	MF	*					*
Gogolla et al. [10, 27]	SP	OCL	MF	*	*				*

Language (ECL), a task-specific model management language, in order to define language-specific algorithms for model matching. The second approach (with partial oracle functions) aims to ensure certain properties of a transformation using the partial oracle functions. It is no need to manually define a whole expected target model within the approach. The works in [8, 28, 7] employ OCL contracts or OCL assertions as partial test oracles to express the expected properties of generated models and to automatically verify them. Contracts and assertions can be represented in the form of visual graph patterns as explained in [7, 31].

Testing is an informal approach for verifying the quality properties of model transformations. Depending on a given test objective, partial oracle functions aim to check whether the functional behavior of a transformation system fulfills such following properties: (1) *confluence*, *applicability* and *termination* which are called general properties; (2) *correctness* including syntactic and semantics correctness (for both information preservation and behavior preservation) and the completeness which are called specific properties. A specific property is often specific to a certain transformation specification. The analysis of general properties such as termination, determinism, rule independence, rule applicability, or reachability of system states requires to perform on a set of given transformation rules. This task is out of the scope of this paper.

The works in [28, 15, 7, 17, 1] propose to verify the syntactical correctness of transformations using test oracles captured from the target model's contracts. To check the source-target correspondence property [29], also known as the information preservation property, current approaches often employ source-target contracts represented either by OCL conditions [8, 7] or graph patterns [36] to consistently specify input test conditions and output test oracles. This work focuses on analyzing the impact of constraints used for test model generation on different transformation properties.

Test adequacy criteria measure the quality of a test suite regarding to several objectives. Test adequacy criteria help define testing goals to be achieved. In transformation testing, test adequacy criteria can be based on how well the test basis (e.g., the input metamodel and the transformation specification) is covered by the test models, or how effective the oracle functions are to identify synthetic bugs (so-called mutants) injected into the under-test transformation. As shown in the two last columns of Table 1, coverage-based approaches propose to measure the effectiveness of a black-box testing approach by evaluating how the input/output metamodels and/or the transformation specification are covered by the testing technique. Fleurey et al. [1] propose to measure the quality of a set of test models by measuring how much they cover the input metamodel. A measurement technique is defined in terms of class coverage, attribute coverage, and association coverage. The metamodel-coverage or effective metamodel-coverage are also introduced in several other works [2, 21, 5, 22, 6, 9, 39]. The notion of transformation specification coverage is introduced in [7, 8]. Within contract-based specifications, transformation contracts can be analyzed to define test conditions. For example, Guerra et al. [7] take preconditions and invariants as transformation properties and define test criteria that could cover these properties for generating test models. Test criteria could also be defined based on the combination of these properties within a combined testing strategy like t-way testing.

Additionally, mutation analysis approaches aim to measure the effectiveness of test cases based on their ability to detect bugs. Mottu et al. [50] propose exploring mutation analysis for model transformations. They study potential bugs that developers may bring into model transformations to define a set of generic mutation operators for model transformations. The mutation analysis technique is commonly used by current works in literature to effectively show the test case generated by proposed meth-

Table 2: Approaches for oracle function definition (**PO**=> Partial Oracle; **CO** => Complete Oracle; **SC** => Type Correctness/Syntactic Correctness; **IP** => Information Preservation)

Reference	Oracle type	Representing expected outputs	Automated	MT Properties
Guerra et al. [7]	PO	Pattern, OCL	*	SC, IP
Fleurey et al. [1]	PO	OCL	*	
Mottu el al. [31]	CO	Pattern		SC
Wieber et al. [15]	CO	Model		SC
Lano et al. [28]	PO	OCL	*	SC
Hilken et al. [8]	PO	OCL	*	SC
Lin et al. [17]	PO	Model	*	SC, IP
Troya et al. [30]	PO	Model		SC
Orejas et al. [34]	PO	Model		SC

ods [24, 7, 15].

3 Running example

This section motivates our work with the CD2RDBM model transformation between class diagrams (CD) and relational database models (RDBM). This transformation example is introduced in [46]. This paper focuses on its simplified version for common transformation situations as regarded in [25]. Metamodels specifying the input and output modeling spaces of the CD2RDBM transformations are shown in Fig. 1 and Fig. 2, respectively. Requirements of the CD2RDBM transformation contain constraints as restrictions on input/output models and the relationship between pairs of them. At the specification level, the requirements are independent of implementation language and often specified in the form of *transformation contracts*.

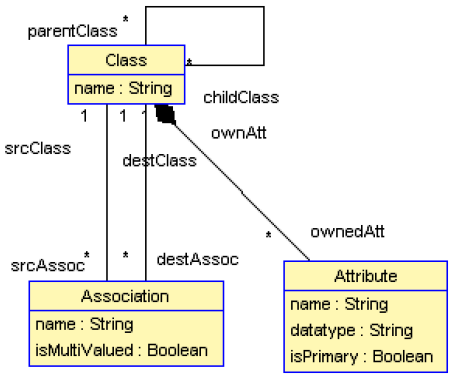


Figure 1: The simplified metamodel of class diagrams.

A transformation contract allows a designer to specify what a transformation does, under which conditions it can be applied to a model, and what its excepted result is. Such information is also helpful for choosing and applying the proper transformation in the context of off-the-shelf transformations. A contract-based model transformation specification typically consists of three sets of constraints corresponding to preconditions, postconditions, and invariants. First, *Preconditions* include constraints defining a set of models, each of which is a candidate as the source model. *Positive preconditions* state the expected properties on valid source models; *Negative preconditions* define source models that fulfill several forbidden properties, i.e., the source ones are invalid. For example, the CD2RDBM transformation includes the following precondition constraints:

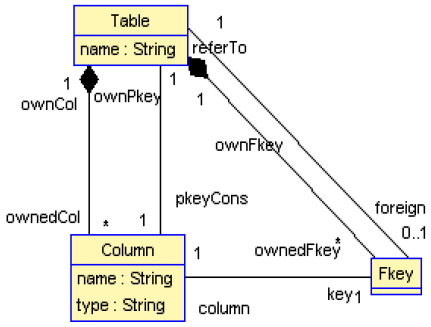


Figure 2: The simplified metamodel of relational database schema.

- A class does not inherit itself;
- The name of a class is unique;
- Attributes of a class must have distinctive names
- The child class does not redefine attributes of its parent class;
- The name of an association does not coincide with a class's name.

Second, *Postconditions* define a set of models produced by the transformation: *Positive postconditions* state the expected properties of valid target models; *Negative postconditions* define target models that satisfy several forbidden properties, i.e., the target ones are invalid. For example,

the CD2RDBM transformation includes the following post-conditions:

– A table name is unique;

– Two columns of a table must have distinct names;

– A table cannot have more than one primary key column.

Third, *Invariants* specify the correspondence between pair of source and target models, denoted by $ps \implies pt$. A positive (negative) invariant has the expressing structure: If the source model satisfies the property ps then the target model (does not) satisfies the property pt. As discussed in [26, 27, 7], the structure of each transformation rule also can be represented by a positive invariant that must hold between the source and target models to satisfy the transformation definition. The CD2RDBM transformation specification contains the following negative invariants:

– If the CD model has two classes with an inheritance relationship, the corresponding RDBM model could not have two distinction tables mapping to these classes.

– If the CD model has two mutually inherited classes, then the corresponding RDMB model could not only have the mapping table to the parent class while there is no mapping table to the child class.

– If the CD model has a class, the corresponding RDBM model could not have two distinction tables mapping to this class.

In addition, the CD2RDBM transformation has six mapping rules that define how a CD model is mapped to a corresponding RDBM model:

– A class must be mapped to a same-name table;

– The name and data type of a non-primary attribute coincides with the ones of a corresponding column;

– A primary attribute is mapped to a column played as the primary key;

– A multi-valued aggregation and association between two classes is mapped to a new associative table that relates the two corresponding tables;

– An aggregation/association relationship between two classes is characterized by a single-valued end and a multi-valued end (0..*, 1..*) is mapped to a foreign key that relates two corresponding tables;

– A child and its parent class are mapped to the same table.

Testing is required to find out if a model transformation is implemented and executed as expected for all possible inputs, or if there are bugs in the transformation leading to unintended output models for certain input models [29]. This model transformation can be realized using different transformation implementation languages. To test the quality of a model transformation captured from multiple restriction domains, a black-box testing approach is often employed.

Since exhausting testing is impossible, testing criteria are proposed to select representative test models to achieve the source metamodel coverage and the specification coverage.

Depending on the test objective, either the positive testing strategy or the negative testing strategy will be used to navigate the test case design and test execution process. The analysis of information on a test basis allows testers to determine test conditions in both negative testing and positive testing strategies.

4 Overview of the approach

Figure 3 overviews our approach to testing model transformations. The basic idea is to synthesize test models based on an integration of multiple restriction domains.

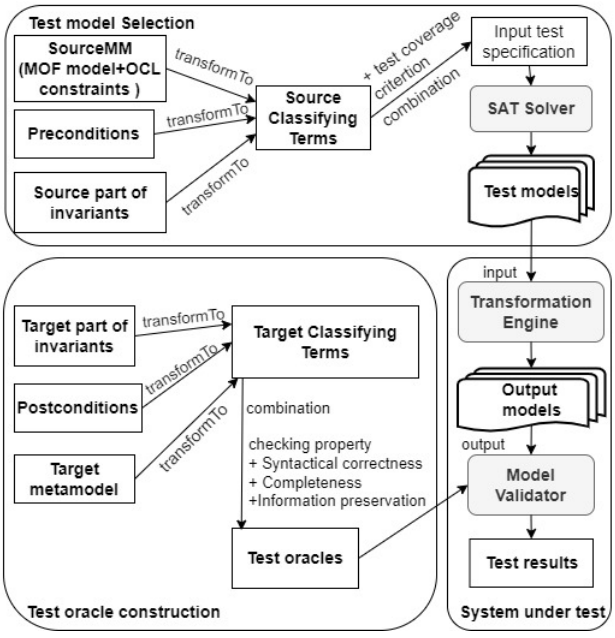


Figure 3: An integration of multiple restriction domains for model transformation testing.

First, the partitioning technique is employed to define test models that cover the source metamodel. The partitioning criteria that are either restrictions on the source metamodel or contracts of the transformation specification are expressed in the form of boolean OCL expressions, referred to as so-called classifying terms (CTs) [8]. In this way, the underlying conditions which are used in characterizing test models also can be flexibly combined to generate effective test suites.

Second, test criteria could be defined for both positive and negative testing strategies. To generate test models that

satisfy a test criterion, input test conditions captured from each restriction domain are expressed by classifying terms. The classifying terms are then combined and taken as the input of constraint solvers, including the SAT solver, in order to automatically generate test models.

Finally, to check test oracles, classifying terms derived from the target metamodel and the transformation specification are defined to ensure that (1) the output model conforms to the target metamodel, (2) the output model satisfies the postcondition, and (3) the output model must also comply with invariants that describe the transformation relationship between valid or invalid pairs of source and target models. Such test output evaluation conditions are then combined to evaluate expected properties on the output model using model validator tools, including OCL tools like USE.

5 Synthesizing test models from restriction domains

Test model selection involves finding valid and invalid input models within positive and negative testing strategies, respectively. Test models are generated by synthesizing models from different restriction sources. This section explains combining knowledge sources to generate valid and invalid models within the positive and negative test strategies based on the proposed covered criteria.

5.1 Metamodel coverage

In MDE, a metamodel is often represented in the form of a UML class diagram with the key meta-concepts of MOF [37] including classes, attributes, generalization, and associations. Therefore, test models, that conform to the source metamodel, can be defined by an equivalence partitioning on the class diagram [38] for the source metamodel with the two following criteria [1]:

- *AEM (Association End Multiplicities):* For each association end, each representative multiplicity must be covered. For instance, if an association end has the multiplicity [0..*], then it should be instantiated with the multiplicity 0, 1, and N (N is greater than 1).
- *CA (Class Attribute):* For each attribute, each representative value must be covered. For instance, representative values of a boolean attribute are $true$ and $false$ that define two corresponding partitions.

These criteria AEM and CA, as illustrated in Table 3, could be expressed in terms of representative values [1, 21, 4]. Representative multiplicity pairs can then be computed for an association by taking the Cartesian product of the possible multiplicities of each of its two ends. The representative values of each attribute can be computed from the typical data types of class attributes such as `Integer`, `String`, and `Boolean`.

Table 3: Representative values for multiplicities

Multiplicity property	Representative values
0..1	0, 1
1	1
0..*	0, 1, [>1]
1..*	1, 2, [>2]
N..*	N, N+1, [>(N+1)]
N..M	N, N+1, M-1, M

Boolean classifying terms (CTs) [39] are used to represent equivalence partitions for test models as follows. For each direction of an association between two classes, the name of the first class, the role name of the second class, and the multiplicity of the association ending at the second class are parameterized by variables $fClass$, $dClassRole$ and $sizeNumber$, respectively. Note that the $sizeNumber$ corresponds to the representative multiplicity value (as depicted in Table 3) at the second class. The parameter $fClass1$ is to define an arbitrary variant of instances of the first class. Using these parameters as input of the following OCL template, boolean CTs are generated from the metamodel.

```
fClass.allInstances -> exists( fClass1 |
  fClass1.dClassRole -> size() = sizeNumber )
```

Figure 4 shows a set of CTs for the simplified class diagram metamodel. Figure 5 demonstrates the partition analysis based on CTs captured from multiplicity values. Test suites with test models generated by the CT set would satisfy the association coverage.

This partitioning approach also includes the restriction on the data type of class attributes. Thus, generated test models could ensure the attribute coverage criterion [44, 2, 4, 3], i.e., each representative value of an attribute must be covered in at least one test model. The following example illustrates how representative values could be defined by analyzing the data range of primitive data types.

- The representative values for `Boolean` attributes are $\{true, false\}$;
- The representative values for `String` attributes: $\{null, '', ' something'\}$;
- The representative values for `Integer` attributes: $\{0, 1, > 1\}$.

The following OCL template is proposed to generate CTs for the attribute coverage criterion.

```
clsName.allInstances -> exists( varCls |
  varCls.attrName = rprValue )
```

In this OCL template, the parameter $attrName$ defines the attribute name, the $clsName$ defines the class name, the $rprValue$ defines the chosen representative value for the attribute data type, and the $varCls$ is to define an arbitrary variant of instances of the class. Figure 6 demonstrates the

```
1  Class.allInstances->exists(c1|c1.parentClass->size()=0)
2  Class.allInstances->exists(c1|c1.parentClass->size()=1)
3  Class.allInstances->exists(c1|c1.childClass->size()=0)
4  Class.allInstances->exists(c1|c1.childClass->size()=1)
5  Class.allInstances->exists(c1|c1.childClass->size()>1)
6  Class.allInstances->exists(c1|c1.ownedAtt->size()=0)
7  Class.allInstances->exists(c1|c1.ownedAtt->size()=1)
8  Class.allInstances->exists(c1|c1.ownedAtt->size()>1)
9  Attribute.allInstances->exists(a1|a1.owningAtt->size()=1)
10 Class.allInstances->exists(c1|c1.srcAssoc->size()=0)
11 Class.allInstances->exists(c1|c1.srcAssoc->size()=1)
12 Class.allInstances->exists(c1|c1.srcAssoc->size()>1)
13 Class.allInstances->exists(c1|c1.destAssoc->size()=0)
14 Class.allInstances->exists(c1|c1.destAssoc->size()=1)
15 Class.allInstances->exists(c1|c1.destAssoc->size()>1)
16 Association.allInstances->exists(a1|a1.srcClass->size()=1)
17 Association.allInstances->exists(a1|a1.destClass->size()=1)
```

Figure 4: CTs coverage representative values of association's multiplicities.

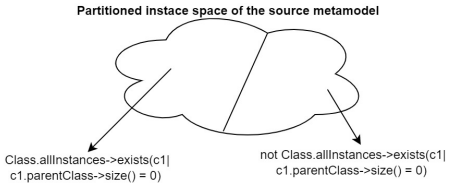


Figure 5: Partition analysis based on CTs captured from multiplicity values.

partition analysis based on CTs captured from the attribute's data type.

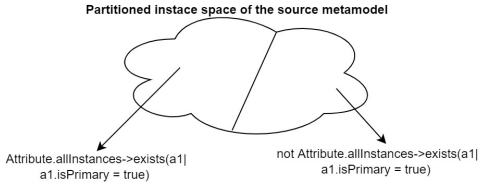


Figure 6: Partition analysis based on CTs captured from attribute's datatype.

There are two basic approaches to select representative values of equivalence partitions. The first, default partitioning chooses representative values using the boundary value analysis or the data random generation. The second, knowledge-based partitioning, representative values are provided by domain experts for various test objectives. This technique also allows the tester to flexibly adjust the configuration to narrow the searching space of constraint solving for a better test model generation.

Figure 7 shows a configuration file defined by the domain expert for the CD2RDBM transformation. Figure 8 shows classifying terms that partition source models based on the properties of the class *Class*.

5.2 Transformation specification coverage

A contract-based specification of a model transformation, brings benefits for debugging, testing, and, more generally, quality assurance. The partition analysis technique can be applied to contracts of a transformation specification to generate the test models that cover the transformation specification's requirements as regarded in [8, 36, 7]. This section explains a partition analysis technique based on classifier terms for model transformations. First, the underlying transformation will be captured by our TC4MT specification language [40]. The language TC4MT employs typed graph patterns in the form of a UML class added with OCL constraints to express transformation contracts. Transformation contracts can be either positive or negative.

Figure 9 shows a negative precondition specified in the TC4MT language. The precondition states that the CD2RDBM transformation rejects any input model in which a child class redefines an attribute of the parent class. Figure 10 shows a negative postcondition for the generated RDBM models. The example postcondition states a restriction on the output models that the column names of a table must be distinct.

Invariants within a transformation contract state how certain structures of an input model should be transformed. An invariant often consists of a source graph, a target graph, and an optional corresponding graph to connect them. A positive invariant that holds on a pair of source and target models would ensure there exists a target graph for each given source graph. With negative invariants, such a target graph should not be found from the target model domain. Figures 11 and 12 show a positive invariant and a negative invariant of the CD2RDBM transformation, respectively.

Considering each input condition on the input modeling space as a testing property, representative values of the property are defined for a testing partition. Then, graph patterns of representing input conditions are translated into boolean OCL expressions using the template as illustrated

```
------ Class'configuration information
Class = Set {c1,c2,c3,c4,c5,c6, c7,c8,c9, c10}
Class_name = Set{'','Person','Employee', 'Project', 'Department'}
Class_childClass = Set{0,1,2}
Class_parentClass = Set{0,1}
Class_ownedAtt = Set{0,1,2}
Class_srcAssoc = Set{0,1,2}
Class_destAssoc = Set{0,1,2}

------Attribute's configuration information
Attribute = Set {a1,a2,a3,a4,a5,a6, a7, a8, a9, a10}
Attribute_name = Set {'', 'eID', 'name', 'fullname', 'pID'}
Attribute_type = Set{'Integer', 'String', 'Boolean', 'Datetime'}
Attribute_isPrimary = Set{true, false}
Attribute_ownAtt = Set{1}

------Association's configuration information
Association = Set {ass1,ass2, ass3, ass4, ass5, ass6, ass7, ass8, ass9, ass10}
Association_name = Set{'', 'workFor', 'employee'}
Association_isMultiValued = Set{true, false}
Association_srcClass = Set{1}
Association_destClass = Set{1}
```

Figure 7: The configuration file for constructing source CTs.

in Fig. 13.

To translate graph patterns into OCL constraints, this schema will iterate over all objects of each contract (lines 2,3). In the case of negative contracts, i.e., the attribute $status$ of all objects equals to -2, the negation operator **not** appears at the first line of the schema. The function $conditions$ is to check a constraint on the underlying objects and their properties. If there exist two objects oi and oj with the same type ($type_{oi} = type_{oj}$) then the condition $oi <> oj$ will be added. The association between two objects will be translated into a corresponding condition, either $oi.role_j \rightarrow includes(oj)$ or $oj.role_i \rightarrow includes(oi)$. The condition function omits the checking of attributes with *undefined* value. Other OCL constraints of the graph pattern will be included in the function **conditions**. Figure 14 shows an OCL condition translated from the precondition shown in Fig. 9: *There does not exist any redefined attribute in the child class*.

A boolean OCL expression can be assigned to one of two values $\{true, false\}$ to specify a corresponding equivalence partition of the input model set. Models that violate a negative precondition will belong to an invalid equivalence partition of the input model set, while the other models will belong to the remaining partition of the input model set. Figure 15 illustrates the result of the partition analysis on preconditions.

Similarly, postcondition contracts as well as invariant contracts could also be translated into boolean OCL expressions to partition the output model set. These OCL expressions will be taken as OCL assertions playing the oracle function to verify actual output models.

6 Generating test models in different test strategies

Model transformation testing aims to ensure a transformation fulfills its requirements (i.e., validation testing) and to discover defects in the transformation (i.e., defect testing). For a certain test objective, the tester would follow a suitable test strategy. This section explains how test models are generated in such different test strategies.

Figure 16 depicts the workflow for test model generation. First, a test basic, including a transformation specification and a configuration of the test model domain, is analyzed and translated into boolean OCL expressions as classifying terms [8] to define partitioning information sets. Second, depending on different testing strategies, partitioning information sets and test criteria describe how the partitions are combined and selected to design test cases. Here, composite partitions are built according to certain specification coverage criteria. The test conditions in both test strategies are defined by combining single partitions using the relational operators $\{and, or, not\}$. Finally, these partition combinations are then taken as the input of an SAT solver [41] to automatically generate test models. For a particular OCL condition, the solver might not find any valid model since the given scope is too narrow, or there is inconsistency in the specification. In such cases, the search scope can be extended interactively by adjusting the solver parameters.

There are two main test strategies for model transformations: (1) A positive testing strategy aims to ensure correctness. This strategy focuses on generating valid input models. The tester could combine restriction domains cor-

```
Class.allInstances->exists(c1|c1.name='')
Class.allInstances->exists(c1|c1.name='Person')
Class.allInstances->exists(c1|c1.name='Employee')
Class.allInstances->exists(c1|c1.name='Project')
Class.allInstances->exists(c1|c1.name='Department')
Class.allInstances->exists(c1|c1.childClass->size()=0)
Class.allInstances->exists(c1|c1.childClass->size()=1)
Class.allInstances->exists(c1|c1.childClass->size()=2)
Class.allInstances->exists(c1|c1.parentClass->size()=0)
Class.allInstances->exists(c1|c1.parentClass->size()=1)
Class.allInstances->exists(c1|c1.ownedAtt->size()=0)
Class.allInstances->exists(c1|c1.ownedAtt->size()=1)
Class.allInstances->exists(c1|c1.ownedAtt->size()=2)
Class.allInstances->exists(c1|c1.srcAssoc->size()=0)
Class.allInstances->exists(c1|c1.srcAssoc->size()=1)
Class.allInstances->exists(c1|c1.srcAssoc->size()=2)
Class.allInstances->exists(c1|c1.destAssoc->size()=0)
Class.allInstances->exists(c1|c1.destAssoc->size()=1)
Class.allInstances->exists(c1|c1.destAssoc->size()=2)
```

Figure 8: Some source CTs generated from the partition analysis on the class $Class$.

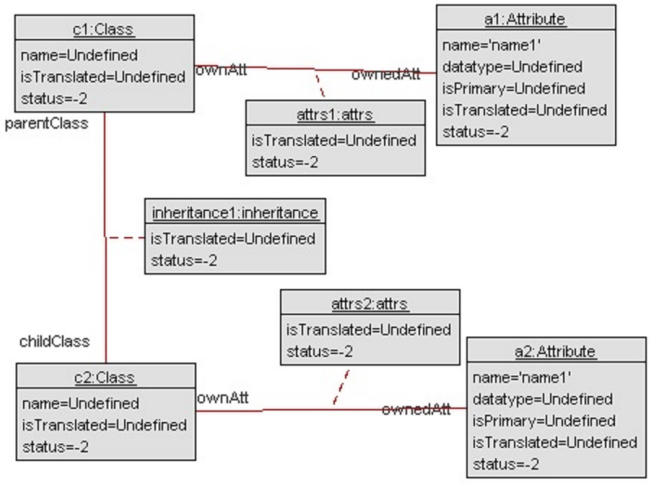


Figure 9: A negative precondition of the CD2RDBM transformation.

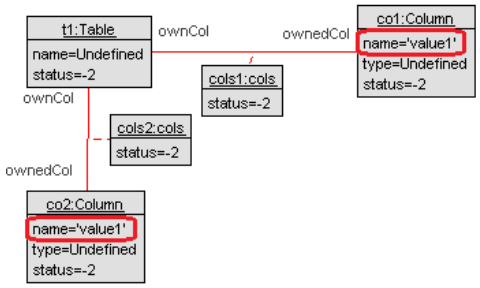


Figure 10: A negative postcondition of the CD2RDBM transformation.

responding to different aspects of the correctness property (including syntax correctness, semantic correctness, information preservation, and behavior preservation) to select relevant input models together with OCL assertions. (2) A negative testing strategy is applied to ensure safety and reliability. The strategy focuses on generating invalid input models so that transformation's defects might be detected.

6.1 Negative testing

Negative testing ensures that a model transformation can gracefully handle invalid input or unexpected execution scenarios. An input model is invalid if it violates at least one negative precondition. The equivalence partition technique is applied to preconditions of the transformation to identify various invalid partitions of input models.

To illustrate the negative testing approach, testers focus on the following typical situation. From a given negative precondition, two equivalence partitions are defined: a set of invalid input models that violate this precondition ($false$) and a set of the remaining models that fulfill this precondition ($true$). A model of the second set can be valid or invalid due to other remaining constraint conditions. Such a negative test case aims to discover defects when robustness testing. By combining many negative preconditions, a smaller partition of invalid input models would be defined.

To automate the generation of test inputs, a combination strategy is defined that describes how values ($true$ or $false$) for negative preconditions are selected such that the underlying coverage criterion is satisfied. The t-wise coverage criterion tends to be chosen for the negative testing approach. The coverage criterion is satisfied if any value combinations of t parameters, i.e., negative preconditions,

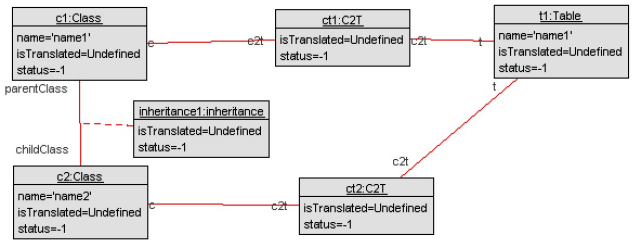


Figure 11: A positive invariant of the CD2RDBM transformation.

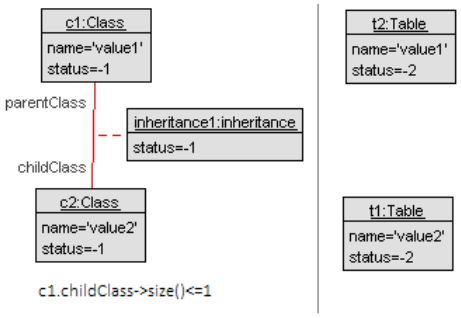


Figure 12: A negative invariant of the CD2RDBM transformation.

in this case, appear in at least one test input. As a special case of this, the following criteria are determined: each choice ($t = 1$), pair-wise ($t = 2$), and exhaustive ($t = n$).

Based on the combinatorial testing with negative test cases for software testing as explained in [42], different levels of specification coverage are defined for the negative test case generation as follows (see Fig. 17 for an illustration).

- **NP coverage:** For each negative precondition (the t-wise coverage with $t = 1$) at least one input model is selected.
- **2NP coverage:** For each negative precondition ($t = 1$) and each pair of negative preconditions ($t = 2$), at least one input model is selected.
- **Combinatorial NP coverage:** For each combination of t negative precondition ($t >= 1$), at least one test input model is selected. For instance, with the case $t = 4$, test input models would be generated for each negative precondition and each combination from 2 to 4 negative preconditions.

```
[not]
type_o1.allInstances->exists(o1|
    type_o2.allInstances->exists(o2|...
        type_on.allInstances->exists(on|
            conditions(o1,o2,...,on))))
```

Figure 13: The OCL schema for the precondition compilation.

```
not
Class.allInstances->exists(c1|
  Class.allInstances->exists(c2|
    Attribute.allInstances->exists(a1|
      Attribute.allInstances->exists(a2|
        c1<>c2 and
          a1<>a2 and
            c1.childClass->includes(c2) and
              c1.ownedAtt->includes(a1) and
                c2.ownedAtt->includes(a2) and
                  a1.name = a2.name))))
```

Figure 14: The OCL expression translated from the negative precondition shown in Figure 9.

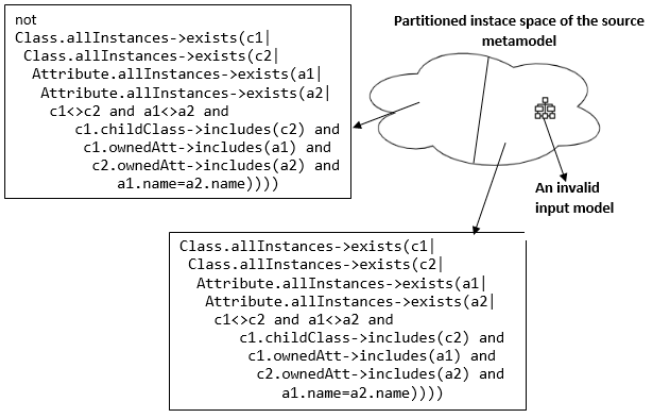


Figure 15: Partition analysis based on CTs captured from preconditions.

Figure 18 shows four test models generated by solving source classifying terms of the CD2RDBM transformation. These classifying terms are defined as a combination of negative preconditions. The first test model (M1) plays the role of a negative test case violating the negative precondition *NoSelfInheritance*. The remaining test models (M2,M3, and M4) are generated by the classifying term, defined by combining the two negative preconditions *NoSelfInheritance* and *NoDuplicateClassName*.

Linking negative test cases with test oracles. Negative testing ensures that a model transformation can gracefully handle invalid input data or unexpected user behavior. The purpose of negative testing is to prevent the system from crashing due to negative inputs and improve its quality and stability. The completeness property requires that the transformation refuses invalid input data and does not contain any incomplete execution. The syntactical correctness property requires that any output model produced from an invalid input model needs to be invalid, i.e., it violates at least one negative postcondition. The completeness of a transformation could be checked by performing negative

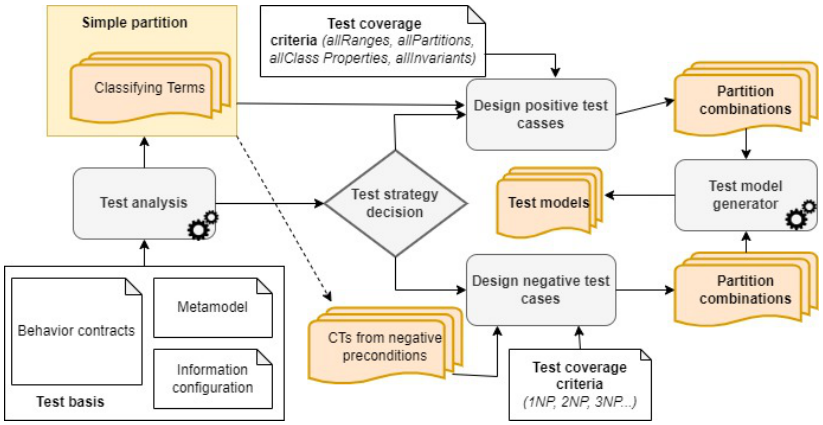


Figure 16: A test model generation process in different test strategies.

a) NP coverage

TC_id	P1	P2	P3
1	False		
2		False	
3			False

b) 2NP coverage

TC_id	P1	P2	P3
1	False	-	-
2	-	False	-
3	-	-	False
4	False	False	-
5	False	True	-
6	True	False	
7	False	-	False
8	False	-	True
9	True		False
10	-	False	True
11	-	True	False
12		False	False

c) Combinatorial coverage

TC_id	P1	P2	P3
1	False	-	-
2	-	False	-
3	-	-	False
4	False	False	-
5	False	True	-
6	True	False	
7	False	-	False
8	False	-	True
9	True		False
10	-	False	True
11	-	True	False
12		False	False
13	False	False	False
14	False	False	True
15	False	True	False
16	False	True	True
17	True	False	False
18	True	False	True
19	True	True	False

Figure 17: Partition analysis based on CTs captured from preconditions.

test cases and observing manually the execution process. The expected output of the execution is either an invalid input warning or a non-terminating state of the transformation system. Similarly, the syntactical correctness of the transformation also can be checked. The output model is now checked using the oracle function as shown in Fig. 19.

6.2 Positive testing

Positive testing verifies how the application behaves for the positive set of data. In positive transformation testing, single partitions are combined to select valid input models, representing composite partitions of the input model domain. Because valid input models must satisfy without violating any negative preconditions, all classifying terms translated from negative preconditions will be pushed into the SAT solver when solving constraints to generate valid input models.

A strategy to combine partition information in terms of classifying terms is defined to avoid duplication and reduce the number of test models. This strategy also ensures both the source metamodel coverage and the transformation specification coverage. The concept *Range* is denoted to a set of equivalent values of a class property and a*Partition* contains one and more *Range*. The following coverage criteria are proposed for the positive testing approach.

The *allRanges* coverage. The representative value of each range must be implemented in at least one test model. The following examples are model fragments of the metamodel(*MF*).

```
MF{Class.allInstances → exists(c|c.name='')}
MF{Class.allInstances → exists(c|c.name='var')}
MF{Class.allInstances → exists(c|c.childClass →
size()=0)}
```

The *allPartitions* coverage. The set of representative values of each *Partition* must appear in at least one test model. The following example model fragments are generated from this fragmentation criterion.

```
MF{Class.allInstances → exists(c|c.name='') ∧
Class.allInstances → exists(c|c.name='var')}
MF{Class.allInstances → exists(c| c.childClass
→ size()=0) ∧ Class.allInstances → exists(c|
c.childClass → size()=0) ∧ Class.allInstances →
exists(c| c.childClass → size()>1)}
```

A test model based on this coverage criterion can represent more constructs to be tested in the source metamodel than the *allRanges* coverage criterion. If an area is divided into three ranges, the tester can create a test model that corresponds to the three instances of the test model set in the *allRanges* criterion. Therefore, creating a suitable *allPartitions* coverage for a test model set can reduce the test case size while ensuring the metamodel coverage criterion.

The *allClassProperties* coverage. Each value association representing the partition of each class' attribute values must be implemented in at least one test model. The following example fragment models are generated from this

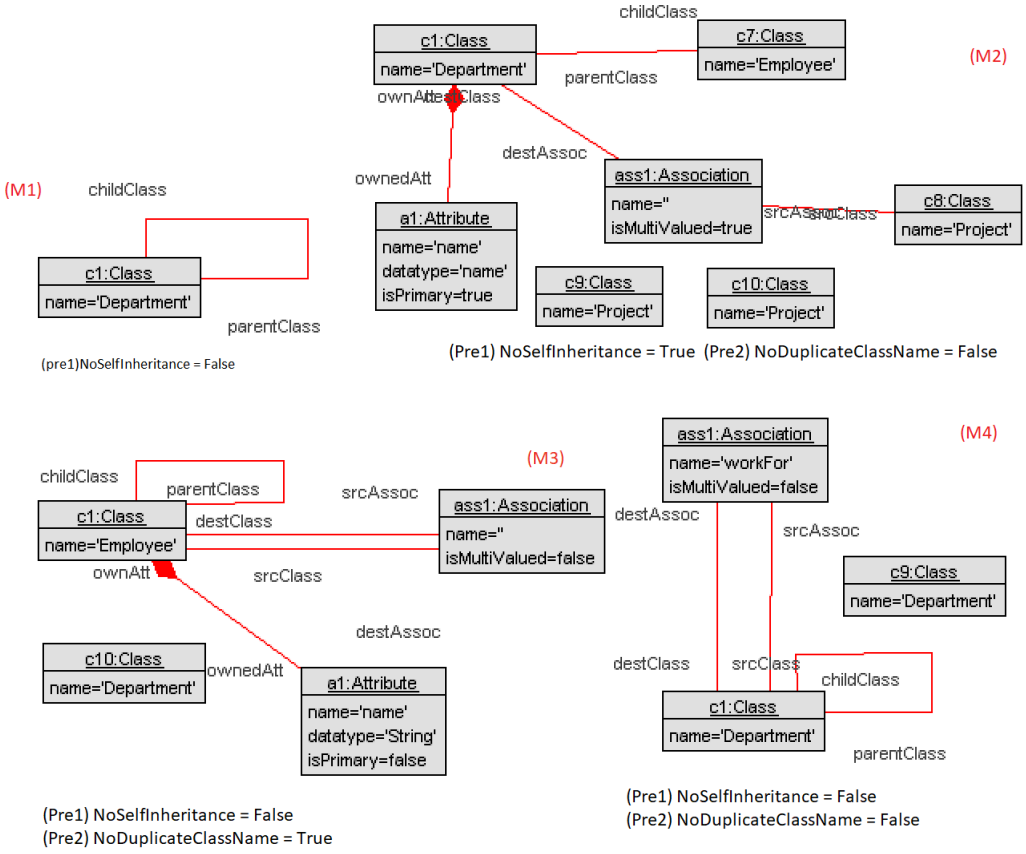


Figure 18: Four negative test cases generated from source CTs.

criterion.

```
MF{Class.allInstances → exists(c|c.name='var' ∧
c.childClass → size()=0 ∧ c.parentClass → size()=0
∧ c.ownedAtt → size()=0 ∧ c.srcAss → size()=0 ∧
c.destAss → size()=0)}
```

```
MF{Attribute.AllInstances → exists(a1| a1.name =
'attname') ∧ Attribute.AllInstances → exists (a1|
a1.datatype = 'atttype') ∧ Attribute.AllInstances →
exists(a1| a1.isPrimary = false)}
```

While the test coverage criteria *allRanges, allPartitions, allClassProperties* allow us to achieve the source metamodel coverage, specification-based test coverage criteria aim at the requirement coverage. A valid input model needs to fulfill all negative preconditions, therefore, test models generated by the positive testing strategy ensure the precondition coverage. In order to achieve the invariant coverage and to navigate the test input model selection, the following test coverage criterion is defined.

The *invariant* coverage. Each source pattern of invariants (consisting of both negative and positive invariants) needs to be implemented in at least one test model.

Positive w.r.t. negative invariants describe valid w.r.t. invalid pairs of source and target models. Therefore, the source graphs as part of an invariant can be used as templates, i.e., positive patterns, to generate test models in the positive testing strategy. The invariant coverage criterion requires that each source graph of invariants (including transformation rules) appears as a restriction on at least one test model. Considering the CD2RDBM transformation, with three negative invariants and six transformation rules, nine test models are required to satisfy the coverage invariant criterion.

Linking positive test cases with test oracles. The completeness property of a model transformation requires any valid input model also to be accepted as the input data and then transformed into an output model. In case all generated output models are valid target models satisfying all negative postconditions, the syntactical correctness property of a model transformation is ensured.

A model transformation is correct only if both input and output models are valid. In other words, the output model must preserve the information as well as the behavior of the input model through the transformation program. The correspondence between source (input) models and target (output) models can be captured by invariants. Therefore, invariants should be effective knowledge sources to check information preservation. Therefore, positive test cases including valid input models should be used as test data for checking the syntactical correctness and information preservation as shown in Fig. 20 and Fig. 21.

```
Function O1
  input:
    mtin: invalid input model
      mt: transformation implementation
  output:
    testResult: Boolean //true(pass), false(fail)
  do:
    mtout:=mt(mtin)
    if(mtout.satisfies(forAll(p)|p: negative postcondition))
      testResult:=false
    else
      testResult:=true
    endif
EndFunction
```

Figure 19: Oracle function for the syntactical correctness in the negative testing strategy.

```
Function O2
  input:
    mtin: valid input model
    mt: transformation implementation
  output:
    testResult: Boolean //true(pass), false (fail)
  do:
    mtout:=mt(mtin)
    if(mtout.satisfies(forAll(p)|p: negative postcondition))
      testResult:=true
    else
      testResult:=false
    endif
EndFunction
```

Figure 20: Oracle function for the syntactical correctness in the positive testing strategy.

```
Function O3
  input:
    inv{ps,pt}: positive invariant,
      ps is the source pattern of the inv,
      pt is the target pattern of the inv
    mtin: valid input model,
      mtin.satisfies(ps}
    mt: transformation implementation
  output:
    testResult: Boolean //true(pass), false (fail)
  do:
    mtout:=mt(mtin)
    if(mtout.satisfies(forAll(p)|p: negative postcondition))
      if(mtout.satisfies(pt}
        testResult:=true
      else
        testResult:=false
      endif
    else
      testResult:=false
    endif
EndFunction
```

Figure 21: Oracle function for the information preservation in the positive testing strategy.

7 Tool support

The USE (UML-based Specification Environment) [43] is the execution environment of the support tool of. The tool includes three main functional components as follows: (1) TC4MT specification tool; (2) Test generator; and (3) Test bench.

As shown in Fig. 22, the first component allows building the TC4MT transformation specification using the USE editor. In this component, the metamodel of a transformation specification is represented by a UML class diagram added with OCL constraints. Patterns of a specification are represented by object diagrams conforming to the metamodel. For example, the class diagram in Fig. 22 shows the metamodel of the CD2RDBM transformation specification while preconditions, postconditions, invariants, and transformation rules are represented by object diagrams created by using the graphical window interface or the scripting language SOIL of the USE editor.

The second component is a USE plugin that performs the specification analysis to define test conditions. Figure 23 shows the GUI of this component. The plugin is activated by loading a triple-type graph. The window *MetamodelAnalysis* (red label **1**) is used to automatically generate source classifying terms from the partition analysis on the source metamodel. An optional configuration file containing information provided by the domain expert can be loaded to increase the expressiveness of the source CTs. The window *SpecificationAnalysis* (red label **2**) is used to load patterns of preconditions, postconditions, and invariants (including transformation rules playing positive invariants) and translate them into CTs using the OCL schemes introduced in Section 5.2.

The last component, as shown in Fig. 24, is also a USE plugin playing the test bench. Test bench-related tasks are to use the Kodkod engine to solve OCL constraints for finding model instances playing test models, invoke the system under test (SUT) with the test models, and pass resulting output models to the oracle function for evaluation. This plugin is activated by loading the source metamodel. It takes as input the specification files of metamodels, transformation definition, source CTs, target CTs, and the Model Validator configuration, all of which are plain text files.

The transformation definition including a set of TGG-based rules is written in the RTL language [45] that can run on the USE tool. The configuration file (including value options for links, attributes, and size of elements) is required to restrict object models. The source CTs file is used to gen-

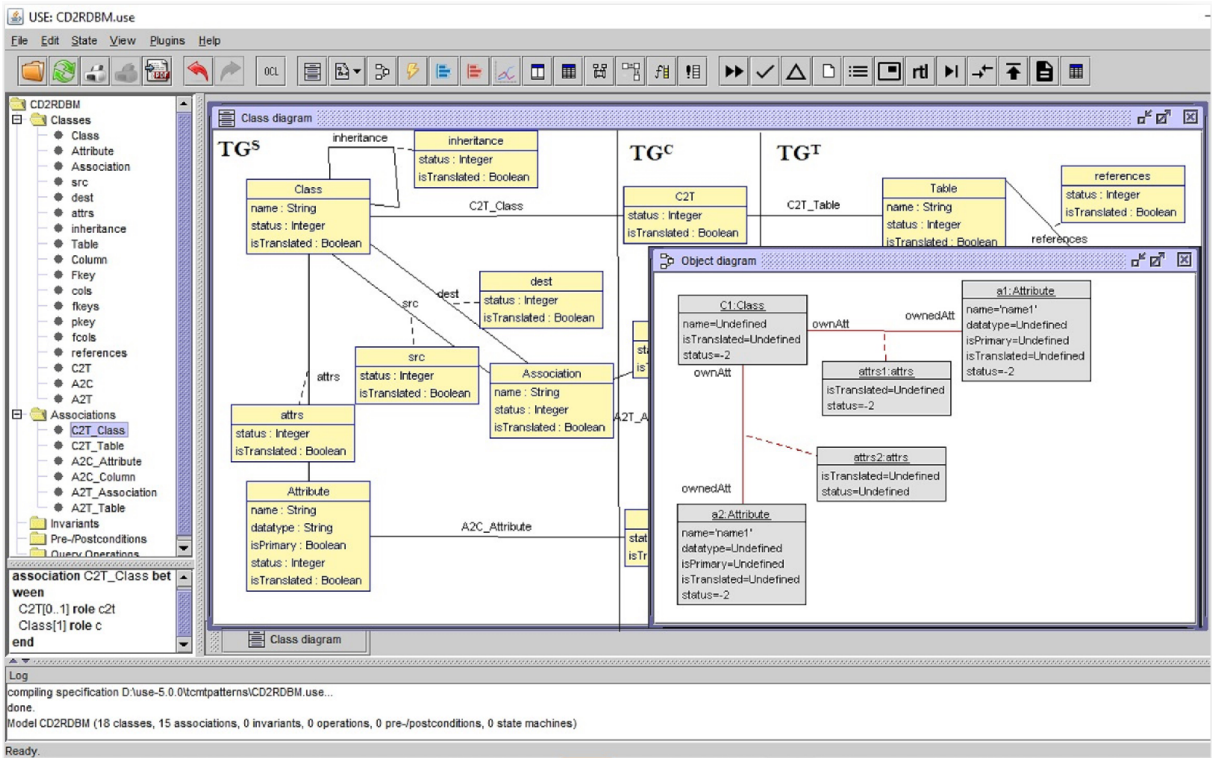


Figure 22: A transformation specification in the language TC4MT.

erate input models of test cases, the target CTs file is used to validate output models. The mapping file contains a list of patterns in the format of ”sourceCTs –> targetCTs”, in which each side specifies a list of expected Boolean values of CTs corresponding to passed test cases. The list of source CTs that are combined to represent the input test specification based on the selected test coverage criteria while corresponding target CTs are combined to represent the expected output property that defines the partial oracle function. The test report obtained from the test bench is shown in Fig. 25.

Finally, as shown in Fig. 25, when executing the test suite, each solution generated by SAT solver will be taken as input for the model transformation. The tool then reports whether the output model satisfies predefined OCL assertions. The partition information of each solution is presented in the panel *Source Classifying Terms*, as shown in Fig. 25. The validation result of the corresponding output model against OCL assertions is shown in the panel *Target Classifying Terms*. The oracle function that is predefined in the mapping file will check whether the test case is passed. The test result will be depicted in the panel *Validation result*. The transformation execution script is shown in the panel *Executed transformations*. The debugging of transformation execution scenarios is performed by invoking each rule application, step by step. The current state of the transformation system after each transformation step could be checked.

8 Experimental results

In this section, several experiments are performed to evaluate the effectiveness of generated input models for detecting transformation failures. The objective of the experiment is to evaluate the error detection ability of the designed test cases in both positive and negative testing strategies.

8.1 Tested setup

For the evaluation, the paper focuses on four model transformations written in the RTL implementation language, the Restricted graph Transformations Language, as proposed in [45]. The purpose of the transformation examples is as follows.

C2R. The CD2RDBM transformation [46] is implemented for the running example which includes six rules, five negative preconditions, three negative invariants, and three negative postconditions;

B2D. The BibTeX2DocBook transformation [47] transforms the BibTeX model into the XML-based format for document composition DocBook. However, in this paper, we are only interested in converting the information about proceedings of conferences presented in BibTeX models into corresponding information presented in DocBook models. The version of the transformation BibTeX2DocBook includes six rules, four

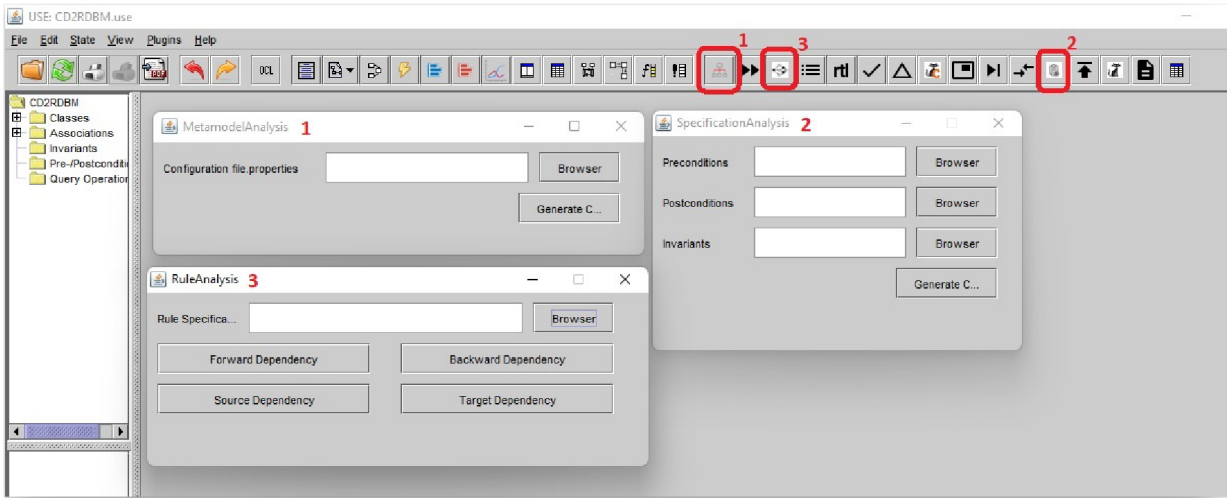


Figure 23: The GUI for the function to analyze transformation specifications.

negative preconditions, two negative invariants, and three negative postconditions;

F2P. The Families2Persons transformation [48] is part of the ATL transformation zoo and was created as part of the "Usine Logicielle" project. This transformation includes four rules, two negative preconditions, two negative postconditions, and no negative invariant;

B2P. The BPMN2PetriNet transformation [49] transforms BPMN models at the Computational Independent Model (CIM) level to PetriNet models at the Platform Independent Model (PIM) level. This model transformation includes twelve rules, five negative preconditions, two negative postconditions, and no negative invariant.

It is important to note here that the specification language TC4MT is independent of transformation implementation platforms. The transformation implementations need to conform to the transformation specification but are not derived from the TC4MT specifications automatically. Generated test suites can be used for verification and validation model transformation implementations using the black-box testing approach.

Table 4 gathers the number of contracts of each transformation example, as well as the size of its input metamodel. In our specification-based testing approach, we focus on negative preconditions, negative postconditions as well as negative invariants. Besides, transformation rules specify expected corresponding mappings between source models and target models so they can be considered as positive invariants of the source-target contracts.

The BPMN2PetriNet transformation is the most complex in terms of the size of specification as well as the input metamodel. The Families2Persons transformation is a simple transformation with few invariants, rules and classes.

8.2 Test suite generation

From the TC4MT specification, a test suite is derived based on each selected coverage criterion. The test suites were automatically generated by using the tool presented in Sect. 7. The numbers in side cell of Table 5 show the number of generated test models corresponding to a particular coverage criterion. In general, the larger the size of the specification is specified, the larger the test size is generated.

8.3 Efficacy of generated test suites

To measure the effectiveness of a test suite and help improve it, the common technique *mutation testing* [50] is employed. In mutation testing, faults are injected into a program to produce erroneous versions of it, which are called mutants. Then each mutant is tested with the test suite. Once the test suite could detect the error, the mutant is killed. Otherwise, the mutant remains alive. The mutant score, which is the number of killed mutants divided by the total number of mutants, gives a measure of the quality of the test suite.

The mutation testing technique is performed as follows. First, mutants of each transformation implementation are created manually by injecting faults by using systematic classification of mutation operators of model transformation regarded in [50].

Navigation. The model has navigated thanks to the relations defined on its metamodel and a set of elements is obtained. Therefore, navigation mutations replace the navigation towards a class with the navigation towards another, remove the last step of a chain of navigation, or add the last step of navigation in a navigation chain.

Filtering. A rule application is usually performed on a limited set of input and output model elements described by the filter conditions. Filtering mutations introduce

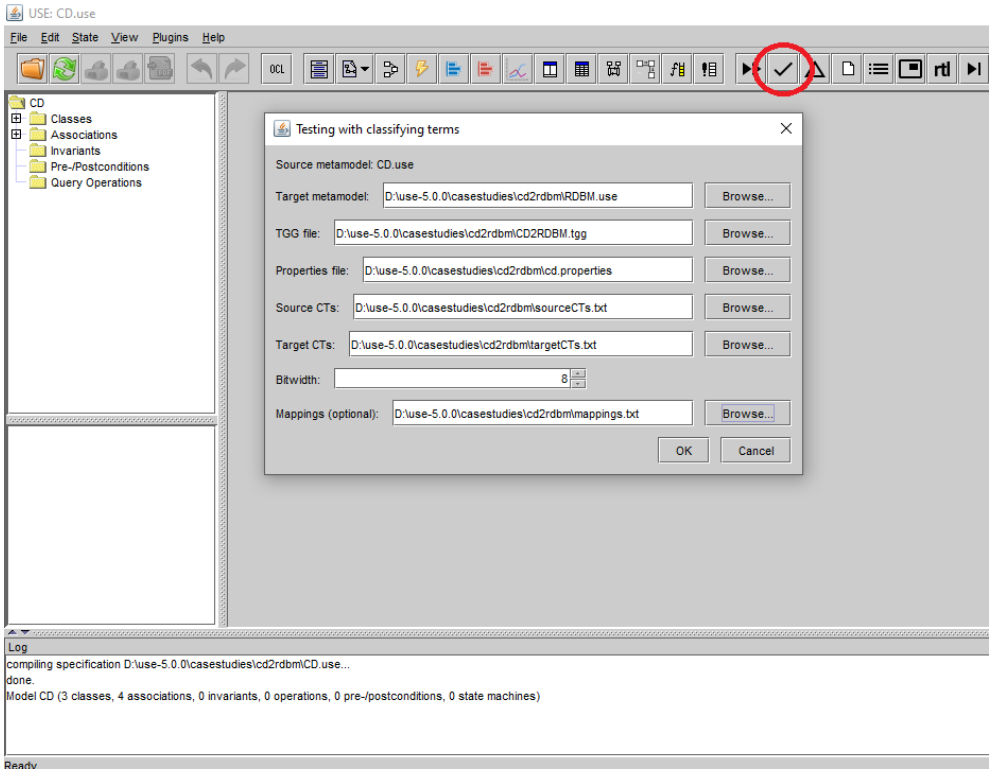


Figure 24: The GUI for the functions: test implementation, execution, and reporting.

Table 4: Test setup

Examples	Specification			Source Metamodel		
	Precond	Postcond	Inv+Rule	Class	Assoc	Inheritance
CD2RDBM	5	3	9	3	4	0
BibTeX2DocBook	4	3	8	5	3	2
Families2Persons	2	2	4	2	4	0
BPMN2PetriNet	5	2	12	15	2	14

disturbances in the filters of a collection, either by modifying the attributes used in the filter or by selecting only some instance types when the collection is defined with a generic class.

Creation. Output model elements are created by the execution of transformation rules. The creation mutations replace the creation of an object with another compatible type, delete the creation of a relation between two objects, or add a useless relation between two objects in the transformation rules of a transformation implementation.

Figure 26 shows an example of mutants. Here, the rule is specified in the RTL transformation language. The injected fault is highlighted in a colored square. In particular, the class $Column$ is related to the class $Table$ by two associations corresponding to the role names *ownPkey* and *ownCol*. This mutant aims to replace $c.ownPkey$ of the column c with the $c.ownCol$ so that the cardinality is modified.

Table 6 shows the mutation operators used to create the mutants in the experiment, which altogether belong to all possible mutation types (navigation, filtering, and creation). Each mutant was created by applying a mutation operator to the original transformation one time. Thus, each cell in the table corresponds to the number of mutants created using a particular mutation operator. The last column in the table summarizes the number of mutants created for each transformation.

Table 7 shows the number of mutants created from each transformation as well as the mutation score of the generated test suites using the negative testing strategy.

Table 8 shows the number of mutants created from each transformation using the positive testing approach.

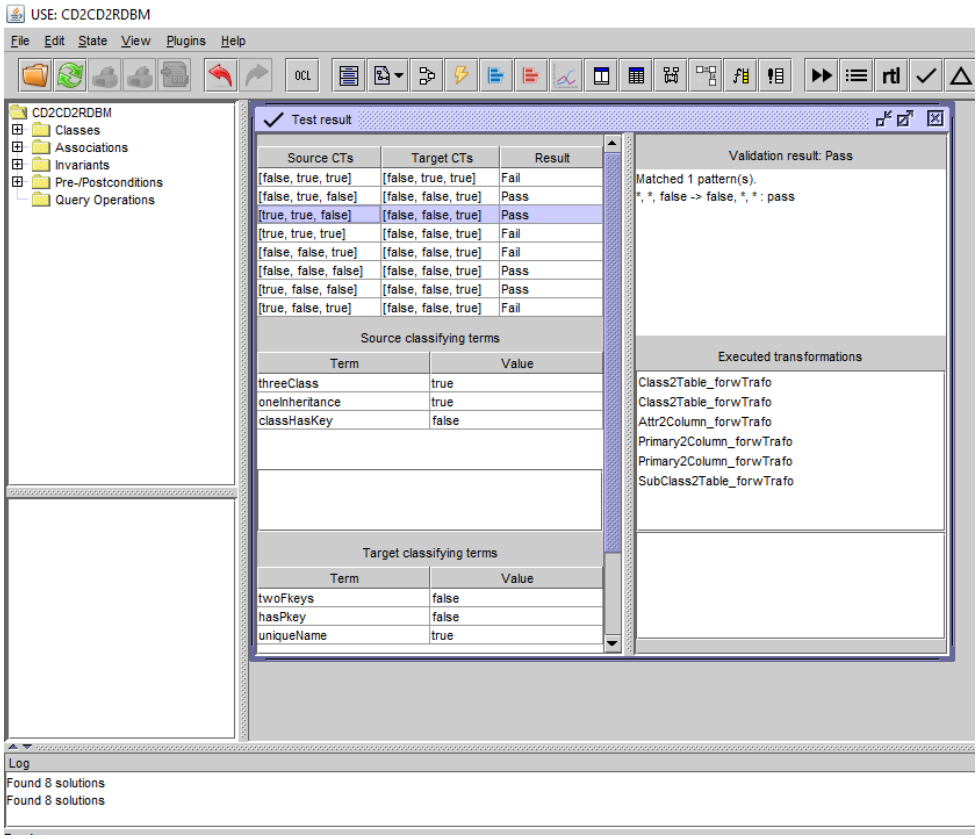


Figure 25: Test report with some partition information.

Table 5: The number of test models in test suites for coverage criteria

	Negative testing			**Positive testing**			
Examples	1NP	2NP	3NP	Ranges	Partitions	ClassProperties	Invariants
CD2RDBM	5	35	63	34	14	3	9
BibTeX2DocBook	4	22	50	24	12	5	8
Families2Persons	2	5	-	10	2	2	4
BPMN2PetriNet	5	35	63	53	30	15	12

Table 6: Number of mutants on the transformations CD2RDBM (C2R), BibTeX2DocBook (B2D), Families2Persons (F2P), and BPMN2PetriNet (B2P)

	Navigation	Filtering	Creation	Total
C2R	9	28	21	58
B2D	6	12	7	25
F2P	12	16	4	32
B2P	13	36	20	69

Table 7: Mutation scores of the generated test suites in the negative testing strategy

	Mutants	1NP	2NP	3NP
C2R	58	0.90	0.90	0.90
B2D	25	0.84	0.84	0.88
F2P	32	0.81	0.81	-
B2P	69	0.62	0.62	0.81

9 Threats to Validity

Although we performed the experiments with utmost care, some underlying parameters potentially threaten the validity of the obtained results:

i) We experimented with common transformation examples that are available in related works. However, we only specify and implement these transformation examples with simplified requirements and particular fragments of input/output metamodels in-

Table 8: Mutation scores of the generated test suites in the positive testing strategy

	Mutants	allRanges	allPartitions	allClassProperties	allInvariants
CD2RDBM	58	0.90	0.90	0.90	0.90
BibTeX2DocBook	25	0.72	0.80	0.80	0.72
Families2Persons	32	0.75	0.75	0.75	0.75
BPMN2PetriNet	69	0.65	0.70	0.70	0.65

```
---r3(n:String,t:String)---single-value-primary-attribute--------
rule PrimaryAtt2Column
checkSource(
    C: Class
    [not Attribute.allInstances->forAll(a|a.isPrimary=true
       and a.owner->includes(C))]
){
    A: Attribute
    (C, A): attrs
    [A.isPrimary = true]
    [A.name<>Undefined]
    [A.datatype<>Undefined]
}checkTarget(
    T: Table
    --original statement
    ---[not Column.allInstances->exists(col|col.ownPkey->includes(T))]
    --mutant with injected fault
    [not Column.allInstances->exists(col|col.ownCol->includes(T))]
){
    Co: Column
    (T, Co): cols
    (T, Co): pkey
}checkCorr(
    (C,T) as (c,t) in c2t:C2T
){
    (A,Co) as (a,c) in a2c:A2C
    A2C:[self.c.name = self.a.name]
    [self.c.type = self.a.datatype]
}end
```

Figure 26: An example for the mutation operator *navigation*.

stead of whole metamodels. For example, in the BibTeX2DocBook transformation, we only work on BibTeX files representing the conference proceedings. Mutation scores of generated test suites generally are dependent on specific factors such as the way to create mutants, the size of test suites, and the quality of the under-test transformation implementation. Therefore, the obtained experimental results only point out the error-detection efficiency of generated test suites for typical semantic faults of transformations. Our mutation-based evaluation method is inapplicable for the other specific faults.

ii) We empirically evaluate transformation examples realized by the RTL bidirectional transformation language designed based on the integration of TGG and OCL [45]. Because of the flexibility of the OCL language, there can be different implementations for the transformation from a specification. Therefore, the number of created mutants for each different implementation does not coincide with each other. Note that the RTL implementation currently is not derived automatically from the TC4MT specification although it also has the TGG-based semantics. Even in the case of the automatically generated implementation, testing such an implementation would affect the evaluation results. This makes RTL implementations more objective assessment.

iii) In the workflow of our proposed approach, several steps are still performed manually and interactively, such as the step to create configuration files containing representative values of partitions, the one to create mutant sets, and the one to select the solver configuration. Therefore, the quality of the tester's work and their decisions should have more or less an impact on the experimental results.

Discussion. As surveyed in Sect. 2, current black-box testing approaches often employ meta-model coverage criteria to ensure that the set of generated input models contains at least one instance of any class or association of the meta-model. They also refer to extreme values for the attributes. However, a limitation of the approaches is that a very large number of test models, including unrelated or duplicated test models, are generated, and the completely generated test model is often not related to the test output evaluation. Several testing approaches focus on contract-based model transformation specification analysis to generate smaller test model sets using the specification-based coverage criteria. An advantage of these approaches is that the test models remain intentional: They are generated for testing a particular combination of transformation requirements so that they can be checked by the oracle function more efficiently.

In this paper, a testing approach is proposed that combines different knowledge sources to generate smaller, more efficient test model suites with different test objectives. This combination reduces test model duplication while still ensuring efficient metamodel coverage and specification coverage. In our approach, the use of environment configuration parameters provided by domain knowledge makes generated test suites more efficient. Test models are designed by using both negative testing and positive testing strategies. The approach allows us to verify further quality properties of a model transformation. Some test oracle functions also are defined for verifying quality properties against appropriate test suites generated by different testing strategies. To show the effectiveness of the generated test cases in detecting common semantic errors, some experiments are performed on different transformation examples as regarded in Sect. 8.

10 Conclusion and future work

This paper proposes a specification-driven testing approach for test data selection. The basic idea is to leverage different sources of knowledge that can be produced during the transformation specification development and to utilize them for automatic generation of test suites. Different sources of knowledge as restriction domains are translated into OCL conditions to facilitate the partitioning testing. The experiments show that boolean OCL expressions could be combined to synthesize test models. Based on the characteristics of knowledge sources and selected testing strategies, input model conditions would be linked with output model assertions to check different quality properties.

The proposal testing framework, named TC4MT, employs SAT solver for finding test models automatically. The TC4MT framework is installed to support automated testing of RTL transformation implementation on USE environments. Several experiments are conducted in which test suites are automatically generated from several transformation specifications. We then measured the efficacy of the generated tests using the mutation analysis. The quality of the generated test set highly relies on how complete a specification is. If a specification only covers part of the transformation requirements, then the generated models may not enable the testing of the underspecified parts.

The performance and scope of test model searching remain a challenge for the proposed approach, we plan to conduct further experiments to improve performance and test coverage.

Acknowledgements

This work has been supported by Vietnam National University, Hanoi under Project No. QG.20.54. We wish to thank the anonymous reviewers for numerous insightful feedback on the first version of this paper.

References

[1] F. Fleurey, J. Steel, and B. Baudry (2004). Validation in Model-Driven Engineering: Testing Model Transformation. *First International Workshop on Model, Design and Validation*, IEEE, pp. 29-40, https://doi.org/10.1109/modeva.2004.1425846

[2] J. Wang, S.K. Kim, and D. Carrington (2006). Verifying metamodel coverage of model transformations. *In Proc. of Australian Software Engineering Conference*, IEEE, pp. 270–282, https://doi.org/10.1109/aswec.2006.55

[3] M. Lamari (2007). Towards an automated test generation for the verification of model transformations. *ACM Symposium on Applied Computing (SAC)*, ACM, pp. 998–1005, https://doi.org/10.1145/1244002.1244220

[4] S. Sen, B. Baudry, and J.M. Mottu (2008). On combining multi-formalism knowledge to select models for model transformation testing. *Software Testing, Verification and Validation (ICST)*, IEEE, pp. 328–337, https://doi.org/10.1109/icst.2008.62

[5] H. Wu, R. Monahan, and J. F. Power (2013). Exploiting Attributed Type Graphs to Generate Metamodel Instances Using an SMT Solver. *In Proc. of TASE*, pp. 175–182, https://doi.org/10.1109/tase.2013.31

[6] S. Jahanbin and B. Zamani (2018). Test Model Generation Using Equivalence Partitioning. *In Proc. of ICCKE*, pp. 98–103, https://doi.org/10.1109/iccke.2018.8566335

[7] E. Guerra (2012). Specification-driven test generation for model transformations. *In Proc. of ICMT* , pp. 40–55, https://doi.org/10.1007/978-3-642-30476-7_3

[8] F. Hilken, M. Gogolla, L. Burgueño, and A. Vallecillo (2018). Testing models and model transformations using classifying terms. *Software and Systems Modeling*, pp. 885-912, https://doi.org/10.1007/s10270-016-0568-3

[9] L. Burgueño, J. Cabot, R. Clarisó, and M. Gogolla (2019). A Systematic Approach to Generate Diverse Instantiations for Conceptual Schemas. *In Proc. of ER*, pp. 513–521, https://doi.org/10.1007/978-3-030-33223-5_42

[10] M. Gogolla, J. Bohling, and M. Richters (2005). Validating UML and OCL models in USE by automatic snapshot generation. *Software and Systems Modeling*, pp. 386–398, https://doi.org/10.1007/s10270-005-0089-y

[11] Stephan Hildebrandt, Leen Lambers, Holger Giese (2013). Complete Specification Coverage in Automatically Generated Conformance Test Cases for TGG Implementations. *In Proc. of ICMT* , pp. 174-188, https://doi.org/10.1007/978-3-642-38883-5_16

[12] J.M. Küster and M. Abd-El-Razik (2006). Validation of Model Transformations - First Experiences Using a White Box Approach. *Models in Software Engineering, Workshops and Symposia at MoDELS 2006*, pp. 193-204, https://doi.org/10.1007/978-3-540-69489-2_24

[13] C.A. Gonzalez and J. Cabot (2012). Atltest: A white-box test generation approach for ATL transformations. *In Proc. of Model Driven Engineering Languages and Systems*, pp. 449–464, https://doi.org/10.1007/978-3-642-33666-9_29

[14] D. Calegari and A. Delgado (2013). Rule Chains Coverage for Testing QVT-Relations Transformations. *In Proc. of the Second Workshop on the Analysis of Model Transformations (AMT 2013)*, pp. 449-464.

[15] M. Wieber, A. Anjorin, and A. Schürr (2014). On the Usage of TGGs for Automated Model Transformation Testing. *Theory and Practice of Model Transformations*, pp. 1-16, https://doi.org/10.1007/978-3-319-08789-4_1

[16] B. Alkhazi, C. Abid, M. Kessentini, D. Leroy, and M. Wimmer (2020). Multi-criteria test cases selection for model transformations. *Autom. Softw. Eng*, 27(1): pp. 91-118, https://doi.org/10.1007/s10515-020-00271-w

[17] Y. Lin, J. Zhang, and J. Gray (2005). A testing framework for model transformations. *In Model Driven Software Development*, pp. 219–236, https://doi.org/10.1007/3-540-28554-7_10

[18] L. Lengyel and H. Charaf (2015). Test-driven verification/validation of model transformations. *Frontiers of Information Technology and Electronic Engineering*, pp. 85-97, https://doi.org/10.1631/fitee.1400111

[19] J.S. Cuadrado (2020). Towards Interactive, Test-driven Development of Model Transformations. *Journal of Object Technology*, 19(1), pp. 1-22, https://doi.org/10.5381/jot.2020.19.3.a18

[20] T. J. Ostrand, M. J. Balcer (1988). The category-partition method for specifying and generating functional tests. *Communications of the ACM*, pp. 676-686, https://doi.org/10.1145/62959.62964

[21] J. Wang, S.K. Kim, and D. Carrington (2008). Automatic generation of test models for model transformations. *In Proc. of Australian Conf. on Software Engineering*, IEEE, pp. 432–440, https://doi.org/10.1109/aswec.2008.4483232

[22] H. Wu (2016). Generating metamodel instances satisfying coverage criteria via SMT solving. *In Proc. of MODELSWARD*, pp. 40–51, https://doi.org/10.5220/0005650000400051

[23] S. Sen, B. Baudry, and J.M. Mottu (2009). Automatic Model Generation Strategies for Model Transformation Testing. *In Proc. of ICMT*, pp. 148–164, https://doi.org/10.1007/978-3-642-02408-5_11

[24] F. Fleurey, B. Baudry, P. Muller, and Y. Le Traon (2009). Qualifying input test data for model transformations. *Software and Systems Modeling*, pp. 185–203, https://doi.org/10.1007/s10270-007-0074-8

[25] H. Ehrig, C. Ermel, U. Golas, and F. Hermann (2015). Graph and Model Transformation - General Framework and Applications. *Monographs in Theoretical Computer Science*, Springer, pp. 5-399, https://doi.org/10.1007/978-3-662-47980-3

[26] J. Cabot, R. Clarisó, E. Guerra, J.D. Lara (2010). Verification and validation of declarative model-to-model transformations through invariants. *J. Syst. Softw*. 83(2), pp. 283-302, https://doi.org/10.1016/j.jss.2009.08.012

[27] A. Vallecillo, M. Gogolla, L. Burgueño, M. Wimmer, and Lars Hamann (2012). Formal Specification and Testing of Model Transformations. *Formal Methods for Model-Driven Engineering*, Springer, pp. 399-437, https://doi.org/10.1007/978-3-642-30982-3_11

[28] K. Lano, S. Fang, and S.K. Rahimi (2020). Model Transformation Specification and Verification. *In Proc. of International Conference on Quality Software*, pp. 45-54, https://doi.org/10.1002/9780470522622.ch14

[29] A.R. Lukman and J. Whittle (2013). A survey of approaches for verifying model transformations. *Software and Systems Modeling*, pp. 1003-1028, https://doi.org/10.1007/s10270-013-0358-0

[30] J. Troya, S. Segura, and A. Ruiz-Cortés (2018). Automated inference of likely metamorphic relations for model transformations. *Journal of Systems and Software (2018)*, pp. 188–208, https://doi.org/10.1016/j.jss.2017.05.043

[31] M. Mottu, S. Sen, M. Tisi, and J. Cabot (2012). Static Analysis of Model Transformations for Effective Test Generation. *In Proc. of ISSRE*, pp. 291–300, https://doi.org/10.1109/issre.2012.7

[32] S. Mazanek and C. Rutetzki (2011). On the importance of model comparison tools for the automatic evaluation of the correctness of model transformations. *In Proc. of IWMCP*, pp. 12–15, https://doi.org/10.1145/2000410.2000413

[33] D. S. Kolovos (2009). Establishing Correspondences between Models with the Epsilon Comparison Language. *In Proc. of ECMDA-FA*, pp. 146–157, https://doi.org/10.1007/978-3-642-02674-4_11

[34] F. Orejas and M. Wirsing (2009). On the specification and verification of model transformations. *In: Palsberg, Semantics and Algebraic Specification*, vol. 5700 of Lecture Notes in Computer Science, pp. 140–161, https://doi.org/10.1007/978-3-642-04164-8_8

[35] L. Addazi, A. Cicchetti, J. D. Rocco, D. D. Ruscio, L. Iovino, and A. Pierantonio (2016). Semantic-based Model Matching with EMFCompare. *In Proc. of ME*, pp. 40–49.

[36] A.C. Carlos and J. Cabot (2014). Test Data Generation for Model Transformations Combining Partition and Constraint Analysis. *In Proc. of International Conference on Model Transformation*, pp. 25-41, https://doi.org/10.1007/978-3-319-08789-4_3

[37] http://www.omg.org/spec/MOF/

[38] A.A. Andrews, R.B. France, S. Ghosh, and G. Craig (2003). Test adequacy criteria for UML design models. *Softw. Test. Verification Reliab*, 13(2), pp. 95-127, https://doi.org/10.1002/stvr.270

[39] L. Burgueño, F. Hilken, A. Vallecillo, and M. Gogolla (2016). Generating effective test suites for model transformations using classifying terms. *In Proc. of PAME/VOLT*, pp. 48–57, https://doi.org/10.1007/s10270-016-0568-3

[40] T.H. Nguyen and D.H. Dang (2021). A Graph Analysis Based Approach for Specification-Driven Testing of Model Transformations. *NAFOSTED Conference on Information and Computer Science*, pp. 224-229, https://doi.org/10.1109/nics54270.2021.9701514

[41] E. Torlak, and D. Jackson (2007). Kodkod: A Relational Model Finder. *In Proc. of International Conference on Tools and Algorithms for Construction and Analysis of Systems*, pp. 632-647, https://doi.org/10.1007/978-3-540-71209-1_49

[42] K. Fögen, H. Lichter (2019). Combinatorial Robustness Testing with Negative Test Cases. *In Proc. of International Conference on Software Quality, Reliability and Security*, pp. 34-45, https://doi.org/10.1109/qrs.2019.00018

[43] M. Gogolla, F. Büttner, and M. Richters (2007). USE: A UML-based specification environment for validating UML and OCL. *Science of Computer Programming*, 69(1), pp. 27-34, https://doi.org/10.1016/j.scico.2007.01.013

[44] E. Brottier, F. Fleurey, J. Steel, B. Baudry, and L.T. Yves (2006). Metamodel-based Test Generation for Model Transformations: an Algorithm and a Tool. *Symposium on Software Reliability Engineering*, IEEE, pp. 85-94, https://doi.org/10.1109/issre.2006.27

[45] D.H. Dang (2009). On integrating triple graph grammars and OCL for model-driven development. *University of Bremen, Ph.D. thesis*, 2009, https://doi.org/10.1007/978-3-642-01648-6_14

[46] https://www.eclipse.org/atl/atlTransformations/

[47] A.G. Domínguez and G. Hinkel (2019). The TTC 2019 Live Case: BibTeX to DocBook. *In Proc. of the 12th Transformation Tool Contest, co-located with the 2019 Software Technologies: Applications and Foundation*, pp. 61-65.

[48] A. Anjorin, T. Buchmann, and B. Westfechtel (2017). The Families to Persons Case. *In Proc. of the 10th Transformation Tool Contest (TTC 2017)*, pp. 27-34.

[49] Z. Li, X. Zhou, Z. Ye (2019). A Formalization Model Transformation Approach on Workflow Automatic Execution from CIM Level to PIM Level. *International Journal of Software Engineering and Knowledge Engineering*, pp. 1179-1217, https://doi.org/10.1142/s0218194019500372

[50] J.M. Mottu, B. Baudry, and Y. L. Traon (2006). Mutation analysis testing for model transformations. *In Proc. of Model Driven Architecture - Foundations and Applications*, 2nd European Conference, pp. 376–390, https://doi.org/10.1007/11787044_28

Implementation of Multiple CNN Architectures to Classify the Sea Coral Images

Zainab N. Nemer[1], Wala'a N. Jasim [2] and Esra'a J. Harfash[1]
[1]College of Computer Science and Information Technology, University of Basrah, Iraq
[2]Department of Pharmacognosy, College of Pharmacy, University of Basrah, Iraq
E-mail: zainab.nemer@uobasrah.edu.iq, walaa.jasim@uobasrah.edu.iq , esra.harfash@uobasrah.edu.iq

Keywords: corals, deep learning, classification images, image processing, coral sea identification CNN, AlexNet, SqueezeNet, GoogLeNet, Inception-v3, coral classification.

Received: September 29, 2022

Image processing and computer vision have a major role in addressing many problems, where images and techniques that are dealt with them contribute greatly to finding solutionsto many topics and in different directions. Classification techniques have a large and important role in this field, through which it is possible to recognize and classify images in a way that helps in solving a specific problem. Among the most prominent models that are distinguished for their ability and accuracy in distinguishing is the CNN model. In this research, we have introduced a system to classify the sea coral images because sea coral and its classes have many benefits in many aspects of our lives. The important thing in this work is to study four CNN architectures model (i.e., AlexNet, SqueezeNet, GoogLeNet/ Inception-v1, google Inception-v3) to determine the accuracy and efficiency of these architectures and determine the best of them with coral image data, and we are shown the details in the research paragraphs. The results showed 83.33% accuracy for AlexNet, 80.85% SqueezeNet, 90.5% GoogLeNet and 93.17% for Inception-v3.

Povzetek: Predstavljena je uporaba arhitektur konvolucijskih nevronskih mrež (CNN) za razvrščanje slik morskih koral.

1 Introduction

There is a growing scientific consensus that earth systems are under unprecedented stress.The human and economic development model developed during the recent industrial revolutions has had a significant impact on our planet. For 10,000 years, the Earth's relative stability has allowed civilizations to flourish. Over time, industrialization has jeopardized this stability. The United Nations Sustainable Development Goals are another lens to see the challenges facing humanity. Six of the 17 goals are directly related to the environment and human influence: combating climate change, wisely using oceans and marine resources, managing forests, combating desertification, islands reverse land degradation and sustainable development. [1]

Effective management depends on ecosystem monitoring, and prompt reporting is necessary to offer timely advice. At the same time, the procedure of gathering underwater data for following the communities which exist under the benthic is greatly aided by digital images. Recent years have seen a tremendous advancement in image recognition technology within artificial intelligence and its various uses in modern society, opening up new technologies and avenues to enhance coral reef monitoring capabilities. Coral reef monitoring is expensive because it requires specialized techniques. Furthermore, due to the remotenessof reefs and diving requirements, long-term data sets are often scattered or spatially constrained. The monitoring method has increased the usage of digital underwater photography over small spatial scales in order to keep costs down [2].In the last year, with the rapid developments in the identification of digital contents, the process of automatic image classification has become the most challenging task in computer vision. In comparison with human vision, the process of comprehending and automatically analyzing images is challenging [3], and as computer vision is a combination of pattern recognition and image processing, the process' output is image understanding [4]. One of the models that have demonstrated excellent performance in computer vision problems, particularly image classification is the Convolutional neural networks CNNs [5]. Currently, CNN has become one of the most attractive methods, and it is now considered as a final factor in many modern, diverse and challenging applications of machine learning applications, for example: ImageNet object detection challenge, image classification, face recognition. A typical CNN Consist of one or more blocks of sampling layers, then it is followed by one or more fully connected layers (FCL) and an output layer, as in Figure (1).

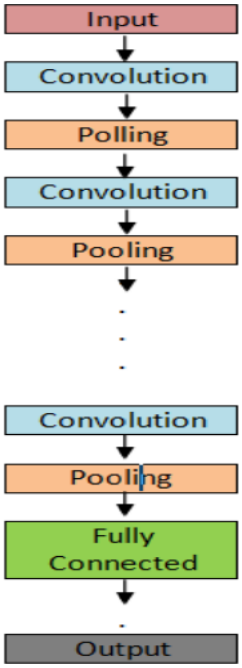


Figure 1: Convolutional Neural Network

The CNN's central parts are the convolutional layer (conv layer). The Images are static typically in nature. That is, the formation of any one part of the image is the same as the formation of any other part. Then, a feature learned in one region may match a similar pattern in another [6]. The CNN model has several architectures, and below we talk about some of them that were used in this work.

The **AlexNet** is a deep CNN. It is used to successfully outperform the classical image object recognition procedures. Rather than a Sigmoid or Tanh, which represented function and were formerly the accepted standards for traditional CNNs, the AlexNet uses ReLu (Rectified Linear Unit) for the non-linear part. ReLu is given by:

f(x) = max (0, x)

Three FCLs are placed after five convolutional layers with reducing filter sizes which are connected (sequentially). AlexNet could quickly down sample the intermediate representations with the use of strided convolutions and max-pooling layers. Vectorized convolutional maps are utilized as inputs to a sequence of two FCLs, as depicted in Figure (2) [7,8].

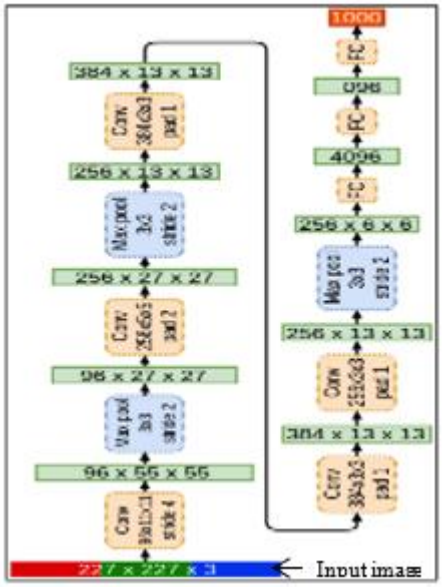


Figure 2: AlexNet arhitecture

SqueezeNet can be defined as one of the CNN architectures that has 50 times less parameters compared to AlexNet while maintaining accuracy on par with AlexNet. Also, this work demonstrated the model's architecture and its application to the ImageNet dataset. The SqueezeNet model employs the following techniques to cut the bulk of parameters: reducing the number of input channels to 3x3 filters, substituting 1x1 filters for 3x3 filters, and down-sampling the network later. Figure (3) shows how the fire module's convolution filters are organized, with a squeeze convolution layer—which has just 1x1 filters—feeding into an expand layer—which has a combination of 3x3 and 1x1 convolution filters [9,10].

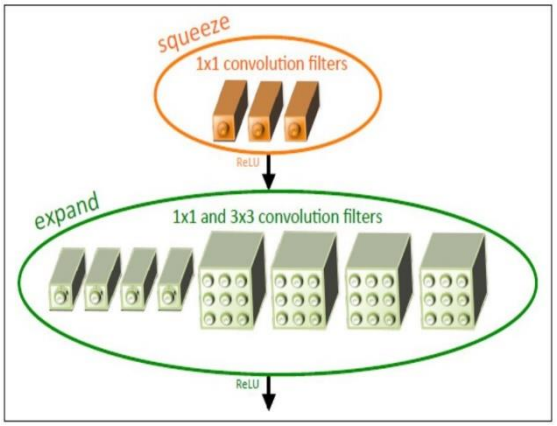


Figure 3: Organization of convolution filters in the fire module.

The GoogLeNet is based on the Inception architecture. It is a system that repeats an inception module. From the network's architecture in Figure 4, it is indicated that there are certain skip connections that, in essence, constitute a mini-module that is replicated across the network. This module was known as an "inception module" by Google. Pooling procedures, spatial convolution, and multiple channel reprojection are all included in each module. Larger convolutional operations (nxn) are split into two convolutional operations with n x1 and n x1 filter sizes. The parameter space is shrunk by two orders of magnitude as a result [11,12,13]

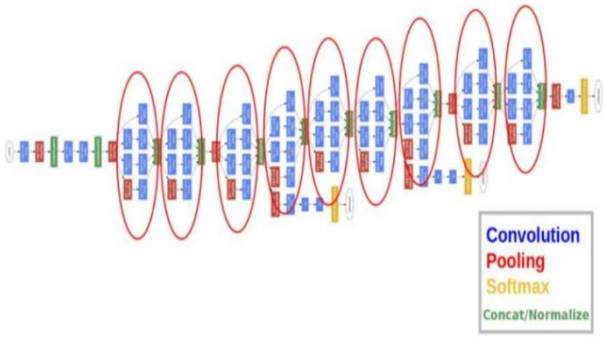


Figure 4: An illustration of the layers of GoogLeNet.

The **Inception-v3** CNN architecture uses Factorized 7 x 7 convolutions, Label Smoothing, and use the auxiliary classifier to transfer label information lower down the network, among other advances (along with using batch normalization for layers in the side head). After that, an FCL is developed on top of the Inception-V3 architecture as a platform for optimizing the process of classification. Convolution layers can learn enough on their own convolution kernel to create the tensor outputs during the model-building process. Additionally, prior to the classification stage, our custom model is concatenated with the individually acquired segmented features. Then it is considered the base of any model because of its capability to get important features that can be utilized in the process of image classification. Figure 5 show the general architecture [14,15].

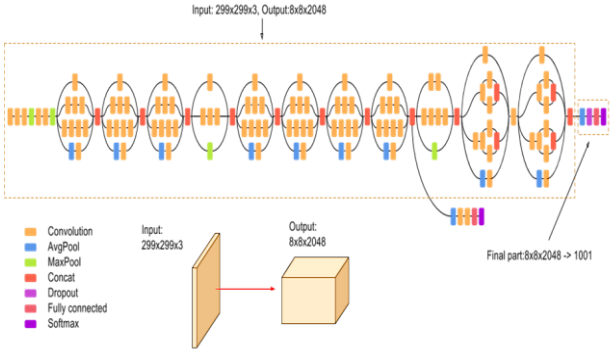


Figure 5: complete architecture of Inception-v3

The four learning transfer architectures have been trained in this study to test their capacity for identifying images of sea coral, and the accuracy results were provided. The rest of this work is structured as follows: Section 1 presents the introduction, section 2 presents the related works, section 3 presents the working system's description, section 4 presents experimental results thoroughly, and section 5 presents the discussions and conclusions.

2 Related work

Convolutional neural network models can be applied to many topics for the purpose of classification. There are many types of CNN models that can be used for each specific topic, and the following is a set of research in this direction.

This study by Sumit Sharan et al. is only based on the challenging but significant Scleractinian (Stony) corals. Further research is done on a suggested method using structural levels like branching corals. The results of the verification show that the testing and training data are nearly identical, demonstrating the capability of the suggested method to accurately predict and learn [16].

S. M. Jaisakthi et al. efforts to automatically recognize and label several types of a benthic substrate using bounding boxes in a given image are introduced as work to monitor coral reefs. In order to recognize and detect various kinds of benthic substrates, an approach based on CNN is given in this research. Since this technique is quicker and more accurate at recognizing objects, they adopted a faster RCNN structure for substrate detection [17].

The classification approach for coral reef images was demonstrated by Zvy Dubinsky et al., and it may be altered to fit other dataset features (number of classes, the size of the dataset, class types, etc.). Also, the study compared several CNN architectures, such as ResNet-50 and VGG-16, and applied transfer learning to the results. There were eleven classes of coral species represented by 5500 images in the ResNet-50 dataset. Here the use of DL is to find out which coral species were most common in the Gulf of Eilat and then link those findings to other ecological factors like water depth or anthropogenic disturbance [18].

Szegedy et al. utilized seven GoogLeNet models in their study. The initialization (and even initial weights, due to oversight) and learning rate policies used for training such models were the same. The main differences between them were the sampling methods they used and the randomness of the input images. The ILSVRC 2014 classification challenge involves placing an image into one of 1000 leaf-node categories in the ImageNet hierarchy. There are around 50,000 validation images, 1.2 million training images, and 100,000 testing images [19].

The purpose of this work, led by Eduardo Tusa and colleagues, is to construct a supervised machine learning-based vision system for coral detections. A bank of Gabor Wavelet filters have been used for extracting texture feature descriptors, and learning classifiers from the OpenCV library have been used to distinguish between non-coral and coral reef. The database

of 621 images (created for this purpose) that depicts Belize's coral reef: Choose the Decision Trees approach since it performs the most quickly and accurately (110 for training the classifiers, 511 for testing the coral detector) [20].

CNNs, a supervised deep learning technique, are used by Mohamed Elsayed Elawady to offer an effective sparse classification for coral species. Additionally, the researchers experiment with cutting-edge underwater image enhancement, color conversion, and color normalization algorithms while computing Phase Congruency (PC), Weber Local Descriptor (WLD), and Zero Component Analysis (ZCA) Whitening to extract shape and texture feature descriptors that are used as supplementary channels (feature- based maps) with the input coral image's basic spatial color channels (spatial-based maps).[21]

The classification of radiography images using 11 CNN architectures (VGG-19, GoogLeNet, SqueezeNet, AlexNet, Inception-v3, ResNet-18, VGG-16, ResNet-50, DenseNet-201, ResNet-101, and Inception-ResNet-v2) is presented by Ananda Ananda et al. With the use of CNNs, two classes—normal and abnormal—of wrist radiographs from the Stanford Musculoskeletal Radiographs (MURA) dataset were identified. Different hyper-parameters against accuracy and Cohen's kappa coefficient were used to compare the architectures. [22]

In order to establish a simpler, more effective, and quicker way to automate the classification of corals, the fundamental analysis was explored in the work of Sumit Sharan and colleagues with the use of approaches like CNN and DL. Only the challenging but significant Scleractinian (Stony) corals are used as a basis for this article. Further research is done on a suggested method using structural levels like branching corals. The results of the verification show that the testing and data are nearly identical, demonstrating the capability of the suggested method to accurately predict and learn [23].

In this article, Nurbaity Sabri and colleagues offer a study that contrasts the leaf recognition abilities of basic CNN and pre-trained models AlexNet and GoogLeNet. The use of such classification models has greatly advanced computer vision. This study uses MalayaKew for detecting leaf recognition performance. GoogLeNet exceeds both standard CNN and Alex Net, achieving a flawless accuracy rate of 100%. Because of the several layers in its architecture, GoogLeNet's processing time is longer than that of the other models [24].

The accuracy of a technique developed by Hopkinson B.M. and colleagues to automatically classify 3D reconstructions of reef sections were evaluated. Locations on 3D reconstruction have been mapped back into the original images to extract various views of the location to produce a 3D classified map. CNNs have been utilized in each method examined for classifying or extracting characteristics from images; however, each method tested differed in the method for combining information from different views of a point into a single classification. Probability averaging, voting and a layer of a learned NN were methods for combining information. [25]-[27]

3 Description of work system

The field of artificial intelligence and computer vision has witnessed during these years tremendous developments with regard to digital image processing and in various disciplines, and this development had a major role in addressing many of the issues that images are mainly involved in solving, including medical, industrial, educational and other issues. In any direction, many factors control the quality of the results, including the size of the amount of data, the method used for processing, and the methods of extracting the final results from the analyzed images. n this research, we turned to treating pictures of sea coral and trying to classify them using the method CNN. The following is a review of the most important steps that were followed in this research to read, treat and classify the sea corals.

There are many types of coral around the world, and there are some species thrive in warm shallow waters and are close to beaches and coasts, and some are located in the depths of the cold, dark sea. So, there are different types of corals in their characteristics, and in general, coral is classified as either hard or soft coral; there are many known types of hard and soft coral. They are easily distinguished because they are similar to plants, live in colonies,and have a distinctive appearance.

For the experiments, we dealt with ten classes of sea corals:(Great Star Coral, Brain Coral, Table Coral, Pillar Coral, Staghorn Coral, Bubble Coral, Sea Pens, Toadstool Coral, Carnation Coral, Gorgonian (Sea Fans). Each class has 50 images. Five of these classes are hard coral, namely:(Great Star Coral, Brain Coral, Table Coral, Pillar Coral, and Staghorn Coral), and the other five are soft coral, namely:(Bubble Coral, Sea Pens, Toadstool Coral, Carnation Coral, Gorgonian). This dataset is compiled accurately and according to accurate specifications of images, and from different sites of the Internet. In the figure 6 samples from each class of the approved coral database.

3.1 The CNN structure of sea coral

In this work, we tested four different CNN networks are:(AlexNet, SqueezeNet, GoogLeNet, and inceptionv3) in order to test the efficiency of each net in terms of its ability to classify sea coral data. The input image is of size 250×250×3 and then cropped to the size that is appropriate for each Net model and what it requires. The following is a description of each network that is used here in this classification problem;

AlexNet: The architecture of AlexNet consists of 25 layers:

- Input data size is [227,227,3]
- There are five Convolutional layers.
- To extract the most appropriate features, there are three of Max-Pooling layers.
- Then two consecutive layers of FCLs,
- Then softmax is used here as the activation layer in

the last network layer for predictions.

- The ReLU activation function, where ReLU is the default activation function,
- Also, the Stochastic gradient descent with momentum (SGD) solver is used.

SqueezeNet: This model is very common in image classification problems because it gives greataccuracy in classification. SqueezeNet architecture consists of 68 layers:

- The input size here is 227x227 x3
- a single convolutional layer of an input and output layer
- Three of 3x3 max Pooling with stride 2
- The Activation Function depends on the ReLU activation function, implemented betweenthe squeeze and expand layers.
- Eight fire modules
- The softmax and the SGD optimizer are used here.

GoogLeNet: GoogLeNet is one of the important models because it is trained faster. The architecture of thisnet consists of 144 layers:

- Input images of size 224x224x3
- Three of 3x3 max Pooling with stride 2
- Nine Inception models
- The ReLU activation function is implemented
- SGD optimizers are used
- Finally, fully connected and softmax

Inceptionv3: In Inception-v3 Architecture, there are 315 layers, and we indicated in this net the Conv comes first, then Batch Norm and ReLU are used after it. The following are some of the properties that applyin this network:

- Input images of size 229x229x3
- four of 3x3 max Pooling with stride 2
- Nine Inception models
- Two grid size reduction
- The ReLU activation function is implemented
- SGD optimizers are used
- The Finally Fully connected
- Then prediction softmax

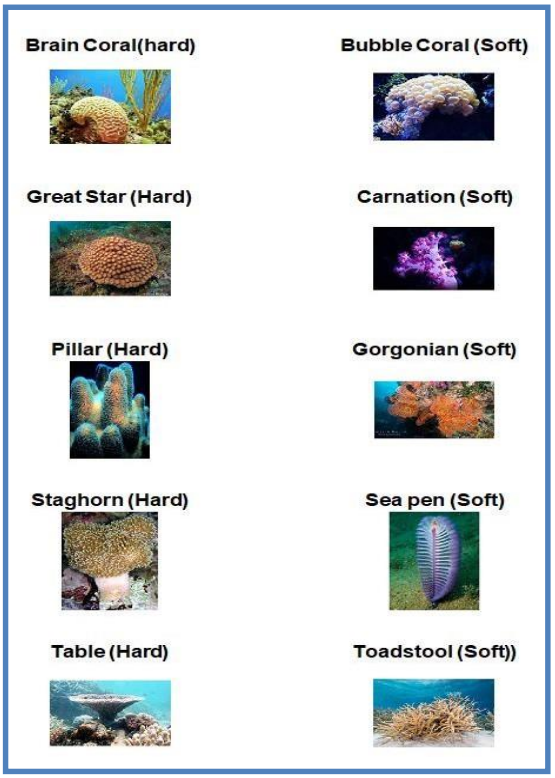


Figure 6: Samples of images of coral dataset

4 Discussion and experimental results

The purpose of implementing several CNN architectures is to know and measure their efficiency in Classification problems, especially in sea coral images, and to determine the most efficient ones. We have trained these nets according to the specifications described above. The CNN architectures:(AlexNet, GoogLeNet, SqueezeNet inceptionv3) are trained on ten classes of coral images. The results obtained with these four CNN models are very encouraging, and the error accuracy of the total results of all ten classes of the coral is shown in Table (1) for 30 epochs.

Table 1: Pretrained deep learning models.

Network	Accuracy validations (Top_1)
AlexNet	83.33
SqueezeNet	80.85
GoogLeNet	90.5
inceptionv3	93.17

These conventional accuracies represent the Top_1, which means the expected answer (the highest probability). All the architectures show important accuracies, but the inception v3 and GoogLeNet achieved higher average accuracy than AlexNet and SqueezeNet. The elapsed time of training of each net is calculated and distributed as in Figure (7). As we can see from the figure, thereis a clear difference in the time that each network spends in the training phase with the stability of the epoch number. Note that inception v3 had the highest training time, although it was the highest accuracy.

With every architecture that is trained, we measure the accuracy of each of ten categories in order to determine the success rate of each type of coral, and the accuracy was measured by relying on Top_5 accuracy (the highest probability answers which should match the expected answer**).** Table (2) shows the details of the accuracy of each class with each architecture.

As is known, the Top_5 method always gives a higher predictor of accuracy, as is evident in Table(2), but from the point of view of careful observation, we find that the Great Star and Sea pens coral are almost better with every architecture.

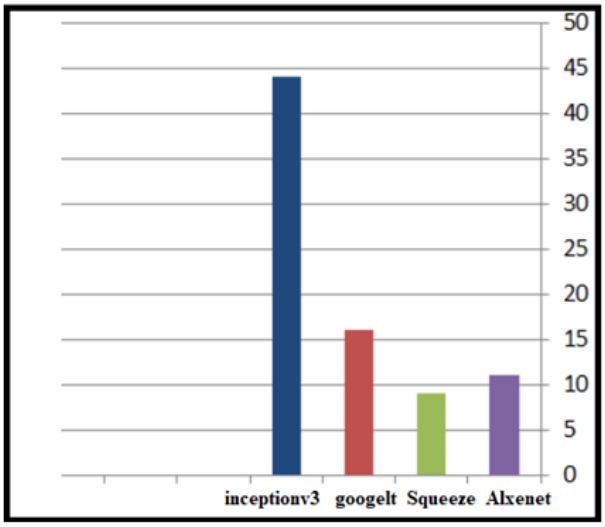


Figure 7: The total training time of architectures.

With every architecture that is trained, we measure the accuracy of each of ten categories in order to determine the success rate of each type of coral, and the accuracy was measured by relying on Top_5 accuracy (the highest probability answers that must match the expected answer**).** Table (2) shows the details of the accuracy of each class with each architecture.

As is known, the Top_5 method always gives a higher predictor of accuracy, as is evident in Table(2), but from the point of view of careful observation, we find that the Great Star and Sea pens coral are almost better with every architecture.

Table 2: Accuracy of each coral class with each network.

Name of coral	AlexNet	Squeeze net	GoogLeNet	inceptionv3
Great Star Coral(hard)	0.9583	0.9916	0.9360	0.9498
Brain Coral (hard)	0.9000	0.9429	0.9513	0.9352
Table Coral (hard)	0.9250	0.9958	0.9017	0.9345
Pillar Coral (hard)	0.9333	0.8941	0.9584	0.9301
Staghorn Coral (hard)	0.8333	0.9428	0.9527	0.9445
Bubble Coral(soft)	0.8667	0.9306	0.9962	0.9299
Sea Pens (soft)	0.9167	0.9958	0.9399	0.9358
*Toadstool Coral (soft)	0.9083	0.8857	0.9656	0.9257
Carnation Coral (soft)	0.8750	0.9875	0.9855	0.9076
Gorgonian(soft)	0.8833	0.9428	0.9236	0.9327

For another test, we trained the architects separately on each type of coral, i.e., hard and soft. This experiment aims to measure each architecture's efficiency in identifying the classesof each type. Table (3) shows the overall results in this case. It is noticeable here that the accuracy error of identifying the classes of each type (hard& soft) was better, but the accuracy of soft type in all the architectures is a certain percentage higher than hard type.

Table 3: Accuracy of each type of coral.

Network	Accuracy of hard coral	Accuracy of soft coral
AlexNet	89.33	90
SqueezeNet	86.83	88.33
GoogLeNet	93.33	95
inceptionv3	96.0	96.67

5 Conclusion

In this research, we have introduced work with Multiple CNN architectures (AlexNet, SqueezeNet, GoogLeNet, inception v3) to classify the sea coral images. The point of view of this workis to know and study the ability of each of architectures in classification problem, especially withthis type of image. In this work, we want to know the possibility of classification of sea coral images by adopting these classification models. We hope at the same time that this work will have a role in clarifying the efficiency and ability of each of these CNN architectures to make it easier to choose any of them according to the data being processed.

Then, what distinguishes this work is the in-depth research to reach results that give a decisions in to directions:first determine the efficiency level of the various CNN archtictures ,each separately ,second

,classifying marine coral and obtaining the best reselts here ,as clarified in the previous paragraph and also in this part. In this system adopts ten types of sea coral, five of which are for the hard coral type and the other five for the soft coral type. Two tests were carried out. In the first test, training of each net(each one separately) on all the ten coral classes, and the final results indicate the high efficiency of allthe architectures in classifying images as Coral, but GoogLeNet and Inception v3 generally recorded better results. The error accuracy with GoogLeNet is (90.5%) and with Inception v3 (93.17%). This is because the GoogLeNet and Inception v3 have distinct architectures in terms of design compared with the rest. They are deeper networks, so their results are generally more accurate.

In the second test, we trained the four nets on each type of coral separately, that is, hard and soft coral, and the results obtained from this test indicated the high efficiency of the four architectures in classification. The GoogLeNet and Inception v3 were also distinguished by relatively higher results than the AlexNet and SqueezeNet, the accuracy of the error with the hard type was (93.33 %) with GoogLeNet and (96%) with Inception v3. And with the soft type was (95%) with GoogLeNet and (96.67%) with Inception v3.

Although the results presented in this paper are very impressive and are sufficient for what we were aiming of this research, some issues may hinder obtaining higher results in this work, including the limited number of images adopted. We believe that if the number of coral images was much greater, the results would have been much higher accuracy. Also, GoogLeNet and Inception v3 take longer time compared to the other models, AlexNet and SqueezeNet, because the number of layers is high in its architecture, especially with Inception v3

Finally, we have tried highlighting the power of CNN models in recognizing the coral images by choosing these four different Architectures. Although all these nets take execution time on the CPU (especially Inception v3), and of course, this time increases with the number of cycles, they are very powerful discrimination models.

References

[1] Celine Herweijer, Dominic Waughray," Harnessing Artificial Intelligence for the Earth", PwC and Stanford Woods Institute for the Environment, January,2018.

[2] Y. Manuel González-Rivero, Oscar Beijbom, Alberto Rodriguez-Ramirez and Dominic E.," Monitoring of Coral Reefs Using Artificial Intelligence: A Feasible and Cost- Effective Approach'', Sensing, Volume 12, Issue 3, p.489. , 2020, https://doi.org/10.3390/rs12030489.

[3] Muthukrishnan Ramprasath, 'Image Classification using Convolutional Neural Networks' **,**International Journal of Pure and Applied Mathematics,Volume 119,No. 17, pp.1307-1319,2018.

[4] Wiley Victor, and Thomas Lucas. "Computer vision and image processing: a paper review." International Journal of Artificial Intelligence Research 2.1,pp. 29-36,2018, https://doi.org/10.29099/ijair.v2i1.42.

[5] Wu, Jianxin. "Introduction to convolutional neural networks.",National Key Lab for Novel Software Technology, Nanjing University, China,Vol. 5, no. 23, p. 495,2017.

[6] F Sultana, A Sufian, P Dutta.,2018, November. Advancements in image classification using convolutional neural network. In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN),,IEEE,pp. 122-129, https://doi.org/10.1109/ICRCICN.2018.8718718

[7] Grm, Klemen, Vitomir Struc, Anais Artiges, Matthieu Caron, and Hazım K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations.",Iet Biometrics vol 7, no. 1,pp. 81-89,2018, https://doi.org/10.1049/iet-bmt.2017.0083

[8] Shadman Q. Salih, Hawre Kh. Abdulla,' Modified AlexNet Convolution Neural Network For Covid-19 Detection Using Chest X-ray Images', Kurdistan Journal of Applied Research (KJAR),Vol. 5,No.1 ,pp. 119-130,2020, https://doi.org/10.24017/covid.14

[9] Forrest N. Iandola, Song Han and Matthew W. Moskewicz, 'squeezenet: alexnet-levelaccuracy with 50x fewer parameters and <0.5mb model size', Computer Vision and Pattern Recognition (cs.CV); Artificial Intelligence, arXiv preprint arXiv:1602.07360, 2016.

[10] Ali Ahmed,' Pre-trained CNNs Models for Content based Image Retrieval',International Journal of Advanced Computer Science and Applications, Vol. 12, No. 7,pp.200-206, 2021,https://doi.org/10.14569/ijacsa.2021.0120723

[11] Gomez-Ríos A., Tabik S., Luengo J., Shihavuddin A.S.M., Krawczyk B. and Herrera F.,' Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation'. Expert Systems with Applications, 118,pp.315-328,2018, https://doi.org/10.1016/j.eswa.2018.10.010

[12] Nur Azida Muhammad, Amelina Ab Nasir and Zaidah Ibrahim,' Evaluation of CNN, AlexNet and GoogLeNet for Fruit Recognition',Indonesian Journal of Electrical Engineering and Computer Science Vol. 12, No. 2,,pp.468-475,2018,https://doi.org/10.11591/IJEECS.V12.I2.PP468-475

[13] Sa Inkyu, Zongyuan Ge, Feras Dayoub, Ben Upcroft, Tristan Perez, and Chris McCool, 'Deepfruits: A fruit detection system using deep neural networks.', *sensors* Vol16, no. 8,p. 1222, 2016,https://doi.org/10.3390/s16081222

[14] Nivrito, A. K. M., Md Wahed, and Rayed Bin, 'Comparative analysis between Inception-v3 and other learning systems using facial expressions detection.', PhD diss., BRAC University, 2016.

[15] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z.' Rethinking the inception architecture for computer vision', In Proceedings of the IEEE conference on computer vision and pattern recognition,(pp. 2818-2826), 2016,https://doi.org/10.1109/CVPR.2016.308.

[16] Sharan, S., Harsh, H., Kininmonth, S., & Mehta, U. (2021). Automated cnn based coral reef classification using image augmentation and deep learning. International Journal of Engineering Intelligent Systems, Vol.29, no.4, pp.253–261,2021.

[17] Jaisakthi S.M., Mirunalini P., Aravindan C., 'Coral Reef Annotation and Localization using Faster R-CNN'. InCLEF (Working Notes),Jan, 2019.

[18] Raphael, A., Dubinsky, Z., Netanyahu, N. S., & Iluz, D.,' Deep Neural Network Analysis for Environmental Study of Coral Reefs in the Gulf of Eilat (Aqaba)'. *Big Data and Cognitive Computing*, *Vol.5*, no.2, pp.19., 2021, https://doi.org/10.3390/BDCC5020019.

[19] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D.,... & Rabinovich, A.,2015,' Going deeper with convolutions.' In *Proceedings of the IEEE conference on computer vision and pattern recognition*,pp.1-9, https://doi.org/10.1109/CVPR.2015.7298594

[20] Tusa Eduardo, Alan Reynolds, David M., Lane, Neil M.,Robertson Hyxia V., and Antonio Bosnjak, 'Implementation of a fast coral detector using a supervised machine learning and gabor wavelet feature descriptors.', In *2014, IEEE Sensor Systems for a Changing Ocean (SSCO).*, pp. 1-6, IEEE, 2014. https://doi.org/10.1109/SSCO.2014.7000371

[21] Elawady M.,' Sparse coral classification using deep convolutional neural networks.', A Thesis Submitted for the Degree of MSc Erasmus Mundus in Vision and Robotics (VIBOT), Department of Computer Architecture and Technology University of Girona, 2014.

[22] Ananda A., Ngan K.H., Karabag C., Ter-Sarkisov A., Alonso E. and Reyes-Aldasoro C.C., 'Classification and visualisation of normal and abnormal radiographs; a comparison between eleven convolutional neural network architectures.', Sensors, Vol. 21,no.16, p.5381, 2021, https://doi.org/10.1101/2021.06.16.21259014

[23] Sharan S., Harsh H., Kininmonth S., & Mehta, U.,' Automated cnn based coral reef classification using image augmentation and deep learning.', International Journal of Engineering Intelligent Systems, Vol. 29,no. 4,pp.253-261,2021,

[24] Sabri N., Aziz Z.A., Ibrahim Z., Rasydan M.A. and Hafiz A.,' Comparing convolution neural network models for leaf recognition.' International Journal of Engineering & Technology,7.3.15,p.141-144,2018, https://doi.org/10.14419/IJET.V7I3.15.17518

[25] Hopkinson B.M., King A.C., Owen D.P., Johnson-Roberson M., Long M.H. and Bhandarkar, SM,,' Automated classification of three-dimensional reconstructions of coral reefs using convolutional neural networks. ',PloS one, Vol.15, no.3,p.e0230671,2020, https://doi.org/10.1371/journal.pone.0230671

[26] Wala'a, N. J., & Rana J. M., (2021). A Survey on Segmentation Techniques for Image Processing, Iraqi Journal for Electrical and Electronic Engineering, vol. 17 , pp. 73-9, https://doi.org/10.37917/ijeee.17.2.10.

[27] Nemer, Z.N., 2022. Hand Gestures Detecting Using Radon And Fan Beam Projection Features. Informatica, 46(5). https://doi.org/10.31449/inf.v46i5.3744

Threat Model and Risk Management for a Smart Home IoT System

Ahmed Redha Mahlous[1*]
[1]Prince Sultan University, KSA, Saudi Arabia
Email: armahlous@psu.edu.sa
[*]Corresponding Author

Keywords: STRIDE, DRED, smart homes, IoT, security risk assessment

Received: November 21, 2022

The emergence of smart homes, driven by the rapid growth and development of technology, has brought numerous benefits to human life, including convenience and improved wellbeing. However, the incorporation of IoT devices into smart homes and their connection to the Internet have created new security and privacy challenges that affect the confidentiality, integrity, and availability of data collected and exchanged by these devices. Such challenges have led to security threats that render IoT devices in smart homes vulnerable to various vector attacks. To provide a comprehensive picture of the security of smart homes, this paper applies the STRIDE [1] threat model to identify potential threats at different layers, namely the IoT device, communication, and application layers. Additionally, a risk-rating security threat model, DREAD, is used to assess the risks of these threats. Finally, this paper proposes a risk mitigation strategy to respond to the rated risks and improve the security of smart home IoT systems. The primary aim of this paper is to enhance the understanding of the various security threats insmart homes and provide a security baseline to enhance the security of smart home IoT systems.

Povzetek: V članku je predstavljena uporaba modela STRIDE na IoT napravah pametnega doma za prepoznavanje potencialnih groženj na različnih ravneh.

1 Introduction

Smart homes or home automation is a term used for homes that have certain devices that sense, control, and regulate the attributes of the house, this might include attributes such as temperature, power consumption, entertainment systems, and might include security features such as camera surveillance and door smart locking.

Smart home devices create a lot of convenience and more control features to homeowners that are extremely attractive to normal homeowners especially when they are at a very competitive price. Benefits include remote control over home features inside or outside the home itself, a decrease in power consumption which creates significant savings for the homeowner, having smart security monitoring which gives a sense of security and privacy for homeowners as well.

The market for smart home and home automation has been increasing dramatically due to the convenience it brings, ease of use and setup, and the decrease of its prices lately due to the huge competition. The global market of home automation is reaching a size of 100 Billion dollars, with more than 250 million homes that use such technologies which represent around 12% of homes worldwide [2].

The competitive nature of such a growing market has also created many flaws that together with many risks and technical issues that are growing as well. Issues and risks may include platform fragmentation [3] which is a term used when many devices with different incompatible software are connected. Lack of technical standards in many of these devices causes more risks that may affect the devices' security and privacy promises. Moreover, the usage of different communication standards also creates many complications when it comes to the security of the systems. And finally, the usage of insecure operating systems such as old versions of android due to the low technical requirement and ease of development imposes huge risks on the security of the systems, with studies that show that more than 80% of android devices that are running are not secure [4], and may have at least one critical vulnerability.

Smart home devices may have many security risks that include easier home intrusions which may happen if the home security system had weak security which allows hackers and thieves to break into the system and disable its feature, or moreover, open the door for them. Also, target targeted attack that targets the smart home device to find and collect data about the user which includes his name, phone number, main email account, password used if it was not encrypted, and maybe their credit cards detail as well. Moreover, a breach of privacy may happen if an attacker had access to previous or even live recordings of any internal camera/microphone which the attacker may use against the victim at any time he wishes as blackmails and more.

Smart home devices have so many kinds of risks due to the amount of point of attack that exists because of their nature, most of them use unprotected communication protocols that are mainly wireless, most of them use unprotected software that controls them, many of them use very weak security policies and controls, and many of them are IoT devices which are connected to the internet which is another point of attack with many kinds of attack as well.

The motivation to write this paper is a rapid increase in smart home devices usage in recent years, and we wanted to explore the different potential threats that can be used against IoT systems in a smart home. The contribution of this study is the result of the risk assessment model which can be used to plan for successful strategy to mitigate risks and contribute to the development of a secure IoT devices for smart home. We believe that it is important to make users and designers of smart home become more aware about the security and privacy breach against such devices.

2 Literature review

Authors in [5] presented a review study of the different face detection approaches in the IoT domain and their application in smart home IoT systems. Authors in [6] surveyed the security of the smart home and the privacy of people living in. They analyzed the security and risks faced by smart home and identified a set of vulnerabilities that can be exploited to gain unauthorized access. The security problems related to the usage of smartphone in smart home was the study of [7]. They listed some problems such as power and Internet malfunction, Software failure, Confidential Data leakage and Eavesdropping attack.

Many studies emphasized on the challenges, risk and difficulties that smart home's owners and designers face in securing their IoT systems [8]. Authors in [9] mentioned the example of the DDOS attack that happened in November 2016 in two buildings in Finland when most of the automated systems controlling thermostats were shutdown.

The data privacy drew the attention of authors in [10] . They highlighted the legal issues related to data privacy and storage in IoT systems in smart home. While authors in [11] tried to fil the gap related to the role of privacy in smart home and address the concerns related to what extend user's concerns for information privacy influenced the intended smart home usage. A multi-theoretical model using Smart PLS 3.2.8 was tested and the derived findings from empirical study emphasize the importance of addressing privacy concerns because they can influence on the intended usage of smart home.

Authors in [12] deducted that user assume that their privacy is protected while using IoT devices but are often unaware about of the potential leak of sensitive information. In another study [13], authors concluded that user's security risk perception has an effect of their intention to use smart home devices, while authors in [14] stated that users convey responsibilities of their privacy protection while using IoT devices to the manufacturers. Authors in [15] provided an overview of users' perception of security while using IoT devices. They developed a model and tested it with multiple linear regression. Using a survey, they concluded that users' awareness about many threats, have an effect on IoT security importance. In the other hand, most of the users do not check their security settings and feel safe while using IoT devices.

The rest of the paper is structured as follows: section 2 literature review, then section 3 presents a scenario and requirements, section 4 reviews Security Objectives of the system. Section 5 presents the risk assessment approach and finally a conclusion is presented in section 6.

3 Scenario and requirements

To understand better the different assets and threats that night exist in a smart home system, we present the following scenario. The surface of the smart home is 200m2 and it consists of two stories building and an attic as shown in Figure 1.

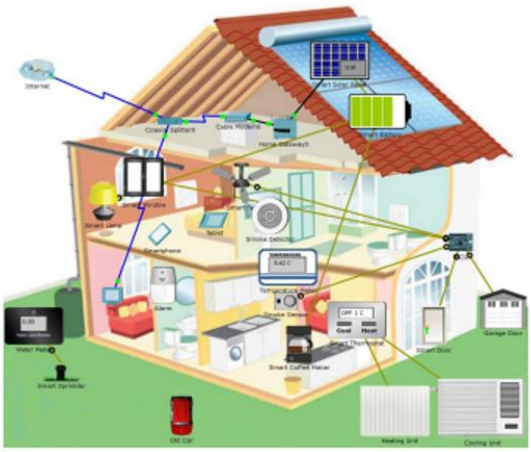


Figure 1: Smart home

The house contains the smart devices (Camera and Smart Door) and some of the controlling devices (tablet), outside of the house, other controlling devices are there such as a Smartphone, all connected to the internet while having an API communicate between the device interface and the user interface.

The user of the smart home system is most of the time away and needs to have the safest house possible. We define the following requirements:

- The user of the smart home system wants to be able to access and monitor the following IoT systems remotely when he is away:
  - o Climate control
  - o Smoke / fire
  - o Temperature issues out of the normal range
  - o Door and window locks
  - o Lawn watering
  - o Local alarm and emergency department messaging
- The user of the smart home system needs to have control over the system locally and through the cloud, which means he should be able to access the controller remotely using his smart phone or locally via a web browser.
- The sensors should send their collected data to the system and different actions should be taken upon the sensor's input. For instance, if the temperature goes above a certain threshold, it probably means that the AC is not working properly, and a notification should be sent to someone without any delay. Also, in case of the presence of a smoke, the smoke detector should sound, and an alert should be sent to the owner as well as to the fire brigade.

- The system should allow the user of the smart home system to change the threshold values that trigger different actuators and events as necessary, either locally or through the mobile app. The triggers and behaviors, data analytics, and remote-control access are all available through a home automation cloud application service that the system will interface with.
- The accounts used to access to the system should be protected by strong passwords.

4 The security objectives of the system

In the smart home we have many different IoT ranging from locks, cameras, and climate controllers to smoke and fire detector and lawn watering, each may have certain logs that store info about their activity or previous recordings. For example, cameras and microphones have previous recordings that are video and voice files. Climate controllers and door locks may have logs about previous activities. All stored info, recordings, and activity logs can be used to a hacker's advantage by doing reconnaissance and data analysis to find more info about the homeowner. All of these kinds of data shall have clear policies regarding their storage and access capabilities to eliminate such risks. Thus, it is imperative to for any system like this to define the associated security needs and objectives. Taking into account the requirements mentioned in section 3 above, we present in Table 1 below the categories, the risk of breaching them and their associated security needs.

Table 1: Categories, risk and security needs.

Category definition	**Risk**	**Security needs**
Identity: access and authorization controls should be in place to document who is accessing the IoT system.	Unauthorized access to the IoT system from stolen credentials.	Each person who accesses the smart home should have a separate Username and password. All access events should be logged in the cloud and retained for a period of time. The actuators in smart home should be controlled by the cloud application, while the IoT systems should have the ability to load their read data to the cloud. In terms of machine to machine (M2M), only allowed machine access is permitted.
Financial: A financial loss due to the system failure should be documented	Substantial cost may incur due to the malfunction or system failure. For instance,	Document the financial losses that could occur due to a failure of the system, system components

	if the climate control fail, the heating or cooling system run unnecessarily.	
Reputation: Customer's reputation might be affected due to the system breach	In the event of security breach, confidential financial information could be stolen such as credit card number. Consequently, the customer's reputation may be damaged.	Document any possible impact on the customer's reputation if the IoT safety/security system is attacked and customer's financial information is stolen.

Privacy and Regulation: Identify any data that could cause privacy concerns for the owner of the smart home system.	Personal financial, health, and other information stored on devices on the network could be stolen	Document the impact of any privacy concerns as well as regulation requirements for this system.
Availability Guarantees: the system should have maximum up time	If the system is down, negative impact to the life of people using the system and damage to the property itself will incur...	No downtime is acceptable
Safety: Ensure the safety of people using the smart home as well the safety of the property	Significant loss to the property and loss of life is the system is compromised.	Document the potential impacts to physical welfare of people and physical damage to equipment and facilities.

5 Risk assessment approach

There are many threats that are documented by known organizations that list vulnerabilities of such devices. Some of the vulnerabilities are reoccurring such as improper authentication techniques. Most vulnerabilities are threats to the confidentiality of the saved data of the smart home system which violates the confidentiality attribute of the CIA model. This attribute specifically is the most important due to the huge amount of privacy concerns and threats generated from such vulnerabilities in this domain.

Cyber attackers today, are becoming more and more clever in launching a cyberattacks against smart home IoT systems due to the existence of many kinds of vulnerabilities that exist in smart home devices, from authentication problems [16] to obtain admin account, insecure storage configuration which allows attackers to gain access [17], and some overflow bugs [18] to listening to open TCP ports to fetch admin passwords [19]. These vulnerabilities cause a potential threat to confidentiality which is the most important aspect of these systems and much more. Thus, it is imperative for smart home designers to be aware of the different threats that might target the smart homes IoT systems.

5.1 Threat model

In this paper we used the STRIDE framework to identify threats, prioritizing and mitigating them. STRIDE is an acronym for each of the threat categories it deals with: Spoofing, Tampering, Repudiation, Information disclosure, Denial of Service, and Elevation of privilege. It was created in 1999 by Microsoft [20].

We created a detailed threat model for the smart home system. For each layer of the attack surface (IoT device Layer, communication layer and application layer), we identified the assets type used in the smart home and the threats corresponding for each STRIDE's category as shown in Table 2, Table 3 and Table 4 respectively.

Table 2: Threat model at the device level.

Threat type	**Asset type**	**Threats**
(S)poofing – can an attacker pretend to be someone he is not, or falsify data?	Sensors	Access to the wireless network through password cracking Man in the middle attack can result in fake data to be injected using bogus devices False sensors can be added to the mart home IoT system
	Actuators	Spoofing the identity of the actuator, thus issuing false control action

Threat type	Asset type	Threats
(T)ampering – can an attacker successfully inject falsified data into the system?	Sensors	Open ports may lead to the access to the smart sensor shell. Theft of sensors. Disconnecting sensors from Power Buffer overflow Sensor stolen or damaged.
	Actuators	Access code theft Theft of actuators. Disconnecting actuators from Power Buffer overflow Actuators stolen or damaged.
(R)epudiation – can a user pretend that a transaction did not happen?	Sensors	-
	Actuators	-
(I)nformation Disclosure – can the device leak confidential data to unauthorized parties?	Sensors	Malware may create false firmware Credentials might be stolen if access to the terminal is achieved. Encryption key and credentials might be disclosed
	Actuators	see above
(D)enial of Service – can the device be shut down or made unavailable maliciously?	Sensors	power source can be disconnected, batteries run out theft or damage
	Actuators	see above
(E)scalation of Privilege – can users get access to privileged resources meant only for admins or superusers?	Sensors	theft of passwords or keys through access to firmware or binaries on the device
	Actuators	see above

Table 3: Threat model at the communication layer.

Threat type	Network or Device	Threats
Spoofing	sensor-actuator network	man-in-the-middle attacks implementation of weak password in 802.1.5.4 security suites
	Wi-Fi Network	Interception and decoding of traffic by a False access point.
	cell phone	same as Wi-Fi using social engineering to trick users to give up passwords
	tablet	man-in-the-middle lost unsecured device allows strangers to access network
	IoT Gateway	weak or default credentials allow access to logs, locally stored sensor data
Tampering	sensor-actuator network	fake device can join network and submit false data lack of message or payload authentication enables false data to be sent on the network
	Wi-Fi Network	wireless protocol security can be hacked, false user joins network and injects false data
	cell phone	-
	tablet	-
	IoT Gateway	wireless protocol security can be hacked, false user joins network and injects false data
Repudiation	sensor-actuator network	time stamping tampered with, damages credibility of logging
	Wi-Fi Network	-
	cell phone	logs of cellular communication not available because of privacy laws
	tablet	-
	IoT Gateway	damage or destruction of any logs on gateway
Denial of Service	sensor-actuator network	rogue device broadcasts on network, keeps devices awake and depletes power wireless signal jamming replay attack ties up network resources or depletes sensor device battery power
	Wi-Fi Network	outdoor APs could be damaged or stolen hacker can use jamming attack which , causes legitimate users' packets to be dropped
	cell phone	-
	tablet	various IP and TCP DoS attacks
	IoT Gateway	ICMP DoS ping attack from outside IP network use of vulnerable UDP services
Escalation of Privilege	sensor-actuator network	interception of weak credentials gains unauthorized access to the network
	Wi-Fi Network	cracked password allows user to gain access weak password on AP allows access to network information and control
	cell phone	weak password on lost or stolen devices allows thieves access to device and configured credentials for other networks

Threat type	Network or Device	Threats
	tablet	same as phone
	IoT Gateway	weak or default passwords

5.2 Applications used in the application layer

Before we define the threats at the application layer, it is essential to know what applications are needed at this layer. The smart home contains a number of applications that help the user to understand what is happening in an IoT system using dashboards and send information about the system.

These applications are accessed through the internet via a web portal and usually are part of a cloud service. Control applications enable interaction with the system, either through direct control of actuators from the application interface, or through software which automates the operation of the system through code that reads sensor values and triggers actuators. We find also embedded applications in some IoT system that can be accessed over the network using HTTP interfaces. Figure 2 shows the applications, how they can be accessed and their purpose.

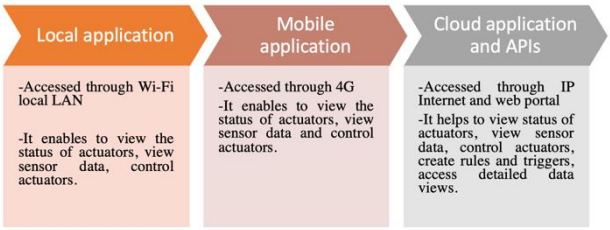


Figure 2: Applications used in smart home.

Table 4: Threat model at the application layer.

Threat type	Application	Threats
Spoofing	local	Wi-Fi man in the middle, packet capture and decryption, false access point enables packet capture
	mobile	stolen phone allows attacker to impersonate legitimate user poorly built mobile apps could use insecure communications mobile apps could steal data or be vulnerable to malware
	cloud	password cracking at web login
Tampering	local	hardcoded credentials, encryption keys, and certificates can be stolen from decompiled firmware, can be used to submit false data
	mobile	unencrypted data may be stored by a mobile app, could be edited
	cloud	unsecured messaging protocols (MQTT) could allow false data to be submitted into the system UPnP opens ports in firewall
Repudiation	local	no logging or transaction tracking
	mobile	insufficient or difficult to access logging of mobile app data

Threat type	Application	Threats
	cloud	insufficient logging, log file corruption or destruction, timestamp tampering logging not available or not configured unreliable logging mechanism
Denial of Service	local	unchanged default passwords enable making IoT devices into bots that work in DDoS attacks
	mobile	multiple failed attempts to log on to device can result in lockout or destroy data
	cloud	repeated brute force attacks intentionally lock out legitimate users DoS attacks against web portal or cloud service
Escalation of Privilege	local	default user accounts and passwords on embedded device apps allow successful logins by unknown users
	mobile	weak or default passwords can enable unauthorized users to access a lost or stolen phone and control the system use on unsecured public Wi-Fi networks may allow hackers to steal credentials and other information
	cloud	SQL injection can provide access to user account information. Weak or default user credentials at web portal allow access to the app across the Internet

5.3 DREAD risk assessment model

The risk assessment model we adopted in our paper is the DREAD [20],[21]. Like the STRIDE model, it was created by Microsoft and it helps rating, comparing and prioritizing the severity of risk presented by each threat that was classified using STRIDE defined earlier in this paper.

DREAD is an acronym that represents the following risk factors: Damage, Reproducibility, Exploitability, Affected users and Discoverability. It averages the scores rated 0-10 for each of risk factor. The higher the number means more serious the risk is, and would be given a higher priority, thus it should be given attention first. Table 5 describes each of the DREAD factors.

Table 5: DREAD factors.

Factor	Definition
Damage	Damage defines the level of damage that could be done to users and the organization if an attack were to succeed.
Reproducibility	Reproducibility is a measure of how easy it is to reproduce a particular attack. For instance, if an attack can be reproduced reliably, it would be rated higher than the one that is statistically unlikely to be exploited or one that cannot be reproduced consistently.
Exploitability	The exploitability of a threat describes how difficult it is to exploit a vulnerability.
Affected users	The affected users risk factor represents percentage of users that will be affected by a particular threat. The greater the number of users who may potentially be affected, the higher this risk factor should be rated.
Discoverability	Discoverability signifies how easy it is to learn about the vulnerability.

In this section, we consider risk metric for some of the relevant threats that have been identified previously. The following assumptions are made:

- All members of the family that live in the home will be affected by any exploit.
- The reproducibility and discoverability metrics always be rated as high (score of 3 for all types of vulnerabilities)
- The Reproducibility and Discoverability are always rated 3.

Table 6: DREAD factor-score

DREAD Factor	Score
Damage	1 = low impact, 3 = high impact
Reproducibility	always 3 - easy
Exploitability	1 = difficult, 3 = easy
Affected Users	1 = few, 3 = many
Discoverability	always 3 - easy

Based on the scoring described in Table 6, a grade is assigned to some of the previously discovered threats from each layer as shown in Table 7.

Table 7: Threat grade.

Attack Surface and Threat	D	R	E	A	D	Total
physical device - firmware can be decompiled and file system and files inspected for credentials or keys	2	3	1	3	3	12
physical device - power source can be disconnected, batteries run out	3	3	3	3	3	15
physical device - data can be faked by bogus devices or injected by man in the middle attacks	1	3	1	3	3	11
communications - lack of message or payload authentication enables false data to be sent on the network	1	3	1	3	3	11
communications - ICMP DoS ping attack from outside IP network	2	3	2	3	3	13
application - unchanged default passwords enables making IoT devices into bots that work in DDoS attacks	1	3	1	3	3	11
application - weak or default passwords can enable unauthorized users to access a lost or stolen phone and control the system	3	3	3	3	3	15

Once the scoring is defined, we put the risks in order by the highest to lowest DREAD metric and estimate the likelihood that the risk will occur. The score of the likelihood is given1 for unlikely and 3 for very likely as shown in Table 8.

Table 8: Threat likelihood score.

Attack Surface and Threat	Total	Likelihood
physical device - power source can be disconnected, batteries run out	15	2
application - weak or default passwords can enable unauthorized users to access a lost or stolen phone and control the system	15	2

communications - ICMP DoS ping attack from outside IP network	13	1
physical device - firmware can be decompiled and file system and files inspected for credentials or keys	12	1
physical device - data can be faked by bogus devices or injected by man in the middle attacks	11	1
communications - lack of message or payload authentication enables false data to be sent on the network	11	1
application - unchanged default passwords enables making IoT devices into bots that work in DDoS attacks	11	3

5.4 Risk response for the rated risks

Once we have identified, categorized, and prioritized the threats to smart home, we provide approaches that document how we want to respond to the threat. As a response to a security risk, we can tolerate the risk, transfer the risk to another party, treat the risk, or terminate the risk as shown in the Figure 3. The detection of threats has value only if there are available responses. Plans for the responses to various attacks should be made in advance. Table 9 is the result of applying one of the responses to the identified threats.

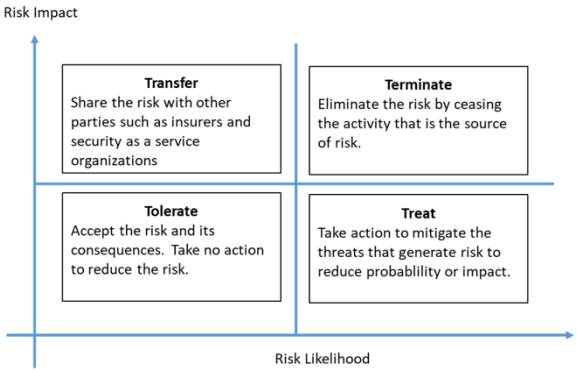


Figure 3: Risk treatment

Table 9: Risk response

Threat	Risk Response
physical device - power source can be disconnected, batteries run out	Treat
application - weak or default passwords can enable unauthorized users to access a lost or stolen phone and control the system	Treat
communications - ICMP ping DoS attack from outside IP network	Tolerate
physical device - firmware can be decompiled and file system and files inspected for credentials or keys	Tolerate
physical device - data can be faked by bogus devices or injected by man in the middle attacks	Tolerate
communications - lack of message or payload authentication enables false data to be sent on the network	Tolerate
application - unchanged default passwords enables making IoT devices into bots that work in DDoS attacks	Treat

5.5 Risk mitigation strategies

Finally, any risks that have been identified with a "treat" response need to be mitigated. Table 10 shows a sample of mitigation strategy for the concerned threats.

Table 10: Mitigation strategy

Threat	Risk Response	Mitigation Strategy
physical device - power source can be disconnected, batteries run out	Treat	because this is a home installation, everyone who lives in the home can be informed that the IoT devices should not be unplugged. For any devices that are on battery, establish a regular day to replace the batteries during the year.
application - weak or default passwords can enable unauthorized users to access a lost or stolen phone and control the system	Treat	Use strong passwords. Inform anyone who has the controller phone app to use strong passwords to protect access to the phone to prevent someone from taking control of the actuators in the house or stealing other information if the phone has been lost.
application - unchanged default passwords enable making IoT devices into bots that work in DDoS attacks	Treat	Change any weak or default passwords. In the design and implementation of this system, the company should enforce a policy that these passwords are changed prior to deployment at the customer site.

6 Conclusion

Smart home devices are great, they give a sense of security to homeowners. Yet, they need constant enhancement to their security measures, many types of security threats exist nowadays from so many types of entry ports. These threats can be resolved with a more standardized way of building these devices and giving them well-designed software that was designed with security in mind. With the current devices in the market, we can see that smart home devices are the weakest link in the chain of devices, so more focus should be put into making them more secure.

Acknowledgment

The author would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

References

[1] https://learn.microsoft.com/en-us/previous-versions/commerce server/ee823878(v=cs.20)?redirectedfrom=MSDN

[2] https://www.statista.com/topics/2430/smart-homes/#dossierKeyfigures

[3] https://www.mobileworldlive.com/mwc16-articles/iot-experts-fret-over-fragmentation/

[4] https://www.zdnet.com/article/android-security-a-market-for-lemons-that-leaves-87-percent-insecure/

[5] Fatima, Saman & Aslam, Naila & Tariq, Iqra & Ali, Nouman. (2020). Home Security and Automation Based on Internet of Things: A Comprehensive Review. IOP Conference Series: Materials Science and Engineering. 899. 012011. https://doi.org/10.1088/1757899X/899/1/012011

[6] Mada Albany, Enas Alsahafi, Itidal Alruwili, Salim Elkhediri, A review: Secure Internet of thing System for Smart Houses, Procedia Computer Science, Volume 201, 2022, Pages 437-444, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.03.057.

[7] Karimi, Khaoula, and Salahddine Krit. "Smart Home-Smartphone Systems: Threats, Security Requirements and Open Research Challenges." 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), July 2019. https://doi.org/10.1109/iccsre.2019.8807756.

[8] Arabo, Abdullahi, Ian Brown, and Fadi El-Moussa. "Privacy in the Age of Mobility and Smart Devices in Smart Homes." 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, September 2012. https://doi.org/10.1109/socialcom-passat.2012.108.

[9] Huraj, Ladislav, Marek Šimon, and Tibor Horák. "Resistance of IoT Sensors against DDoS Attack in Smart Home Environment." Sensors 20, no. 18 (September 16, 2020): 5298. https://doi.org/10.3390/s20185298.

[10] Sanchez, Veralia, Carlos Pfeiffer, and Nils-Olav Skeie. "A Review of Smart House Analysis Methods for Assisting Older People Living Alone." Journal of Sensor and Actuator Networks 6, no. 3 (July 21, 2017): 11. https://doi.org/10.3390/jsan6030011.

[11] Guhr, Nadine, Oliver Werth, Philip Peter Hermann Blacha, and Michael H. Breitner. "Privacy Concerns in the Smart Home Context." SN Applied Sciences 2, no. 2 (January 21, 2020). https://doi.org/10.1007/s42452-020-2025-8.

[12] Zheng, Serena, Noah Apthorpe, Marshini Chetty, and Nick Feamster. "User Perceptions of Smart Home IoT Privacy." Proceedings of the ACM on Human-Computer Interaction 2, no. CSCW (November 2018): 1–20. https://doi.org/10.1145/3274469.

[13] Klobas, Jane E., Tanya McGill, and Xuequn Wang. "How Perceived Security Risk Affects Intention to Use Smart Home Devices: A Reasoned Action Explanation." Computers & Security 87 (November 2019): 101571. https://doi.org/10.1016/j.cose.2019.101571.

[14] Haney, J.; Acar, Y.; Furman, S. "It's the Company, the Government, You and I": User Perceptions of Responsibility for Smart Home Privacy and Security. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Online, 11–13 August 2021

[15] Nemec Zlatolas, Lili, Nataša Feher, and Marko Hölbl. "Security Perception of IoT Devices in Smart Homes." Journal of Cybersecurity and Privacy 2, no. 1 (February 14, 2022): 65–74. https://doi.org/10.3390/jcp2010005.

[16] https://www.cvedetails.com/cve/CVE-2018-9162/

[17] https://www.cvedetails.com/cve/CVE-2018-15123/

[18] https://www.cvedetails.com/cve/CVE-2018-20299/

[19] https://www.cvedetails.com/cve/CVE-2017-11634/

[20] https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling

[21] https://learn.microsoft.com/en-us/windows-hardware/drivers/driversecurity/threat-modeling-for-drivers.

Prediction of Heart Disease Using Modified Hybrid Classifier

Rishabh Pipalwa[1], Abhijit Paul[2*], Tamoghna Mukherjee[3*]
[1]Department of Information Technology, Amity University, Kolkata, India.
[2]Department of Information Technology, Amity University, Kolkata, India.
[3]Department of CSE, Amity University Kolkata, India.
Emails: rishabhpipalwa@gmail.com, a_paul84@rediffmail.com, tamoghna.9@gmail.com
*Corresponding author

Keywords: heart disease, cardiovascular disease, clinical diagnostic system, modified hybrid classifier, health care system

Received: July 5, 2021

This paper proposes a Machine Learning or ML-based strategy to accurately identify a possible heart disease patient. Unlike traditional diagnostic systems which are time-consuming and have human error involved to take care of the patient and diagnose the patients. The proposed system identifies whether the patient will face these kinds of diseases in near future or not. The system is developed based on machine learning techniques such as Naive Bayes, XGBoost gradient classifier, support vector machine, and decision tree. Some external factors were also considered which may lead to heart disease in the future. Furthermore, an integrated web application has been developed that alert and gives a user-friendly interface for recognition and prediction. Thirteen diagnostic factors and five environmental factors are analyzed. The proposed diagnosis system attained good precision as compared to previous methods recommended earlier. In addition, the system can easily be implemented in the public domain to spread awareness regarding heart disease, and it also talks about the possibility of heart disease in near future.

Povzetek: Predstavljeni sistem zazna morebitno srčno bolezen iz trinajstih diagnostičnih in petih okoljskih dejavnikov z uporabo algoritmov strojnega učenja.

1 Introduction

Cardiovascular diseases are popularly known as heart disease leading to a heart attack. In 2018 heart attacks killed nearly 17.9 million humans all over the world. Heart disease is found in 3 out of 5 patients in the critical care unit. The complexity of this disease lies in the fact that it suddenly fails the functioning of human and then SOP (Standard Operating Plan) is required; if not provided on time, patients' life is in danger. A proper healthcare system takes time to detect the cause and effectively start the diagnosis whereas our proposed system efficiently and accurately tells the client whether a patient has heart disease or not. The heart has an essential and critical role in the physical body as it is in control of the flow of blood in different parts of the body which helps in adequate oxygen supply and nutritious elements to be supplied to the required part. Any life is dependent totally on a proper flow of blood, in human life heart is the pumping room of blood. Any disturbance in the flow or the function of the heart may lead to death within seconds [1]. According to the World Health Organization, 17,000,000 people die every year among the 3,000,000 who die before the age of 60 from heart disease. In 2019, the percentage of sudden deaths from heart disease ranged from 4% in high-income countries to 42% in low-income countries [2].

When the heart receives limited blood for a longer period of time it is called ischemic heart disease. Search conditions develop over a course of time which can be periodically monitored and cured with the help of expert supervision. There is a time when ischemic heart patients have a heart attack and after that, the chance of survival also reduces as the disease has been developing over a longer period of time and the heart is habituated or accustomed to limited blood flow. For such things, early predictions or alertness help in the long run.

Diagnosis of heart disease is usually done by reading the patient's medical history, the medical examination report, and the evaluation of symptoms associated with a medical doctor. Although the research found in this diagnostic method is less accurate in diagnosing a heart disease patient. In addition, it is expensive, and it is a computerized challenge to analyze [3]. We have proposed a machine-based diagnostic method. In this study, the machine learning prediction model includes naive bayes, support vector machine (SVM), tree decision, and XGBoost gradient classifier. The standard state of these models has been maintained for analysis purposes. Stalog, Hungarian, Switzerland, Long Beach VA, and Cleveland datasets combinedly were used in this article. We have designed a web-based application that accesses the model for general public use.

This article addressed the problem of predicting the possibility of a heart disease using machine learning (ML) techniques. Here standard feature extraction and profound algorithm classifiers appropriate features were

extracted and analyze with expert guidance from medical's experts which gave a good result in the analysis and accuracy of the proposed algorithm. Then it predicts the future possibility of heart disease by understanding the environmental factors and common habits which may lead to heart disease. Finally, all the modules are combined into a single Python-based framework known as a flask for giving the model a front-end part. This web-based application represents heart disease possibilities with simplicity so that any non-technical or layman can easily detect heart disease.

The organization of the following sections is explained below. This article aims to provide a literature review on relevant heart disease factors and their identification techniques in section 2. Section 3 introduces the proposed system model. Section 4 introduces the web-based design of the proposed model. Section 5 includes results and discussion where performance is analyzed and training and testing results are shown.

2 Related work

The researchers in this case analyze several automatic learning algorithm-based diagnosis strategies to find heart illness. The analysis provides a few machine learning-based methods that make it easier to comprehend the suggested approach. Detrano et al. [4] method for classifying heart illness using machine learning approaches produced a precise end result with an accuracy of 77.00 percent. The dataset was utilized to extract features from the system's multi-layer kit architecture. Another researcher, Gudadhe et al. [5], developed a diagnosis method for heart dis-ease labeling utilizing a multi-layer operational design and SVM classifier and achieved a precision of 80.41 percent. The categorization algorithm for the cardiac disease was created by Kahramanli et al. [6] using a neural network and fuzzy logic. The categorization algorithm achieved a precision of 87.40%. An ANN troupe-based method of heart disease detection was developed by Li et al. [7]. In addition to using a numerical measurement method, it achieved 89.01 percent precision. A machine learning-based approach for identifying heart disease was developed by McKinley et al. [8]. The ANN-DBP system, in conjunction with the FS algorithm, proved worthwhile. A professional health diagnosis approach for heart dis-ease identification was advised by Palaniappan et al. [9]. The prognostic ML models Decision Tree (DT), Navies Bayes (NB), and Neural Networks were utilized to improve the system. Decision Tree Algorithms acquired a precision of 80.40 percent, ANN ac-curacy-ness of 88.12 percent, and Navies Bayes attained a precision of 86.12 percent. Olaniyi et al. [10] developed a 3-layer algorithm for heart disease prediction based on neural network technology and achieved 88.89 percent accuracy. A classification scheme for heart disease employing restraint and stringent set procedures was suggested by Liu et al. [11]. The approach has a 92 percent accuracy rate. For the purpose of identifying cardiac illness, Samuel et al. [12] developed an integrated medical aid system based on Fuzzy AHP and an artificial neural network. The performance of the suggested approach in terms of precision achieved is 91%. Cross-machine learning approaches were employed in one of the research publications by Singh et al. [13] designed as a heart disease forecast tool. They also suggested a novel technique for comprehensive characteristic selection from the data for effective ML classifier training and testing. They were noted as having 88.07 percent accuracy. Sequential Backward Selection Technique for Features Selection, a selection and classification algorithm, has been proposed in [14]. The suggested strategy achieved excellent levels of accuracy. Geweid et al.'s analysis of sophisticated Support Vector Machine-based dichotomy optimization algorithms for heart disease identification [15]. Prior attempts to diagnose heart disease had certain limitations, and the results have been compiled to help people better appreciate the significance of our suggested method. Among all the available techniques, many ways are utilized to spot coronary heart disease early on. Reduced accuracy and lengthy computation times in those earlier solutions are major problems, and it's possible that these are related to the use of datasets with the wrong functions. The prediction must be enhanced for increased detection accuracy, and it also needs to develop effective and accurate early detection for better treatment and healing. Ahammad et al. [16] proposed an approach for designing a healthcare social media platform for services for provisioning, consuming, enabling patients to find an alternate source of healthcare advice, and then it builds a collaborative health community for all kinds of people. Gadiparthi et al. [17] proposed a model for predicting ill effects. Here it predicts the effects of human exposure to social networks in the near future. Milioris et al. [18] investigated and implemented a technique to assess health professionals' views on the adoption and value of health information systems and to assess their usage. Jasim et al. [19] implemented CNN based model for building a system to recognize diseases that are happened in citrus.

Considering the significant research gap and difficulty in improving forecast accuracy, new methods are being used in our paper to precisely locate coronary heart disease to address these problems.

3 Proposed system model

The proposed system model has used a Hybrid classifier that refers to the system being a composite mapping of four algorithms (naive bayes, decision tree, support vector machine, and XGBoost gradient classifiers). The mapping referred to the design of the system in an additive form such that the accuracy of the system gets increased and the error rate reduces because too many systems to run against.

3.1 Data set

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository at the following link: https://archive.ics.uci.edu/ml/machine-learning-

databases/heart-disease/. Stalog, Hungarian, Switzerland, Long Beach VA, and Cleveland datasets combinedly used in this article, featuring the following variables with their description. The size of the dataset is 1023. For training purposes, 648 data are used and for testing purposes, 412 data have been used. In all the classifiers same testing and training ratio has been maintained to get the optimal result. The dataset consists of 13 features dataset where one is the output label output level has 2 possibilities one being the presence of heart disease second being the absence of her disease. Table 1 gives the description of 13 features of the dataset with the feature code. Table 2 shows various external factors and its description which may result in future heart disease.

Table 1: Features of data set with their description.

Sl. No.	Feature name	Feature code	Description
1	Age	AGE	Age in Years Avg=54.38
2	Sex	SEX	Male=1, Female=0 Ratio=70:30 (Male to Female)
3	Chest Pain	CPT	Atypical angina=1 Typical angina=2 Asymptomatic=3 Non-anginal pain=4
4	Resting Blood pressure	RBP	In Mm hg
5	Serum Cholesterol	SCH	In Mg/dl
6	Fasting blood sugar>120mg/dl	FBS	True =1 False =0
7	Resting electrocardiogram	RES	Normal=0 ST T=1 Hypertrophy =2
8	Maximum heart rate	MHR	Numeric
9	Exercise induced angina	EIA	Yes =1 No=0
10	Old peak=ST depression induced by exercise relative to rest	OPK	In Numeric
11	The slope of peak Exercise ST Segment	PES	Up sloping=1 Flat =2 Down sloping =3
12	No. of major vessels Colored by fluoroscopy	VCA	(0-3)
13	Thallium Scan	THA	Normal=3 Fixed Defect=6 Reversible Defect=7
14	Label	LB	Patient has heart diseases=1 Heathy Person =0

Table 2: External factors with their description.

Factors	Feature code	Description
Body Mass Index	BMI	True=has higher BMI False=BMI normal
History of diseases	Phist	Yes =factor present No =Factor not present
Family history of diseases	Fhist	Yes =factor present No =Factor not present
Alcohol	Alchol	Yes =factor present No =Factor not present

Four suitable and efficient classifier techniques are described in a gist in Table 3. In Table 4, a comparison of these four models with the proposed model is also shown.

Table 3: Classifier algorithms with their description.

Classifiers	Description
Navies Bayes algorithm	This is used for the classification concerned problem. The training data set is used by the algorithm to compute the value of the conditional probability of a vector for a given class. The conditional probability value is evaluated for each vector, and then the new vector class is evaluated based on its conditional probability.
Support Vector Machine algorithm	This is a supervised learning model with associated learning algorithms that analyze data for classification and regression analysis. The SVM algorithm is mostly used for classification problems because of its excellent performance in various applications.
Decision Tree algorithm	Shape is the just like a tree consisting of a leaf or addition node. A decision tree has internal external nodes linked to each other. The decision-making part of the internal node takes the decisions and informs the child node to visit the next note.
eXtreme Gradient Boosting algorithm	XGBClassifier of gradient boosting algorithm provides a wrapper class to allow models to be treated like a classifier or regressor.

Table 4: Classifier algorithms with their limitation, advantage and accuracy.

Classifiers	Limitation	Advantage	Accuracy in percentage
Heart disease diagnosis using a single machine learning classifier	Accuracies are very low and system errors can occur very easily	Computation is less complex	70 to 80%
Decision tree + SVM	More exaggeration time is required to generate the result	Accuracy is comparatively high	82.01%
SVM + kNN + k-Means	Computationally complex and performance time is very High	accuracy is high	87.4%

Classifiers	Limitation	Advantage	Accuracy in percentage
System based on Navies bayes + Decision tree + ANN	Computationally complex and ANN performance is low	Navies bayes and decision tree achieved high performance in terms of accuracy	84.33%
Random forest+xgboost + Decision tree	Random forest showed less accuracy in comparison to other classifiers	Xgbosst showed high accuracy	88.21%
Navies Bayes + Decision tree + Support vector machine + XGboost	More execution time is required to generate results	Performance is high and accurate. It suggests high performance in extreme situations	98.73%

Heatmap shown in Figure-1 clearly reflects about the variable weightage which helps in understanding the relevance of each variable.

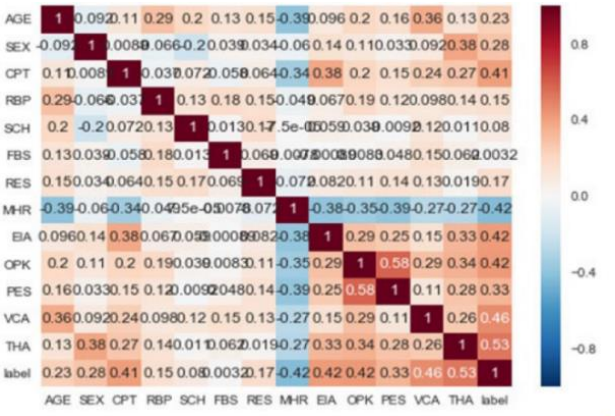


Figure 1: Heatmap of dataset reflecting weightage of each variable.

3.2 The modified classifier used in the model

The modified hybrid classifier is a mechanism to use multiple classifiers with different strategies to get desired output with high accuracy.

3.2.1 Navies bayes

This classifier uses standard arguments in a modified manner to get desired outcomes. Here probability (P) is calculated based on the likelihood of heart disease and class prior probability.

$$\text{P(heart disease)} = \frac{\text{P(Likelihood of heart disease)} * \text{Class prior probability}}{\text{Predictor prior probability}}$$

3.2.2 Decision tree

This classifier is an application for choosing several different extracted features in a modified manner to get desired outcomes.

A function that calculates the degree of randomness, also called entropy is defined as

$f(s) = -P_+ \times \log(P_+) - P_- \times \log(P_-)$ where

P_+ is the probability of the patient having heart disease.

P_- is the probability of the patient not having heart disease.

3.2.3 Support vector machine

This classifier partition training is set into two classes. It maximizes the distance between two parallel hyperplanes of the two classes and minimizes the sum of classification errors. Here f is the optimal function that minimizes total risk.

Min f $= \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m}\rho_i$ where

$\|w\|$ is the distance between two hyperplanes.

ρ is the deviation of misclassified objects.

Here **the** first term of the objective function is structural risk and the second term is an empirical risk.

3.2.4 XGBoost

This classifier uses fast learning through parallel and distributed computing and offers efficient memory usage. It uses begging and boosting. Here boosting technique makes use of trees with fewer splits.

$f_{m-1} + h_m(x) \rightarrow f_m(x)$

'm' denotes the iterations, until residuals have been minimized as much as possible.

'f' is defined to predict the target where f_m is the current model and f_{m-1} is the previous model.

'h' denotes the fit to the residuals from the previous step.

3.2.5 Modified hybrid classifier

Here we have taken 4 independent classifiers based on the accuracy of our dataset. We have experimented and found high accuracy using Naive Bayes, Decision tree, SVM, XGBoost. After this, we propose that each new data will go through all the classifiers and result in individual results.

These results then will be cross-validated with each other to check for any ambiguity. In case of ambiguity, we would go with XGBoost (as it resulted in the highest accuracy on our dataset used). If no ambiguity arises in the result, we will go for result analysis. If the result is positive which means the patient is having heart disease, we display the message accordingly. If it does not have heart disease the possibility of future possibilities and display a message according to the external factor's possibilities.

Strategies used in the modified hybrid classifier are shown below.

Step-1. We would first take the 13 features from the user interface and run them among all 4 classifiers.
Step-2. In case the same value is found (all of them predicting the same disease analysis), we would display the message accordingly.
Step-3. In case of ambiguity (different value from the classifiers), we consider the estimated value given by XGBoost as it resulted in the highest accuracy on our dataset used and display the message of XGBoost on the screen.
Step-4. In case we see that person is not having a heart disease we check with 5 future possibilities variables and display a warning message accordingly.

3.3 Algorithm for the proposed model

Step-1: It starts with the training of the dataset, in which 624 data are trained to each algorithm classifier.
a. There are 4 algorithm classifiers in the model namely Decision tree, Navies Bayes, Support vector machine, and XGboost.
b. A decision tree is a graphical representation for getting all possible solutions to a decision-making situation on a given condition. It follows the supervised learning technique, where internal nodes represent the feature of a data set and branches represent the decision rules and each leaf node represents the possible outcome.
c. Navies Bayesis a probabilistic classifier that predicts the possibilities given by a probability of an object. It applies Bayes law which is based on the probability of a hypothesis with prior knowledge.
d. In the Support vector machine, we plot each data item into a point in the 'n' dimension space where 'n' represents a number of features available in the data set. Then classification is performed on the hyperplane that differentiates the two classes properly.
e. XGboost or extreme gradient boost is an advanced version of gradient boosting classifiers. The major difference lies in the fact XGboost is a regularized model formalized to control overfitting which gives better performance.

Step-2: The extracted feature is computed after training of data set for every algorithm classifier upon which each variable can be used for the model.
Step-3: All the extracted features are sent to 4 different ML algorithms and a resultant output is obtained without any ambiguity.
Step-4: If there is any ambiguity between the four different algorithms the system alerts its reserved feature of taking the most accurate method among all.
Step-5: The resulting output shall be checked with 420 data for testing purposes and the reliability of the model proposed
Step-6: The resulting output is converted into a model which segregates the prediction of heart disease and future possibilities of heart disease.
Step-7: When a user enters new data, it follows a certain pattern to label it into categories.
Step-8: Features are extracted from the new data. Then it is passed to the proposed model.
Step-9: Then a prediction is made about the possibility of having heart disease or not. If a person does not have heart disease at present, then future possibilities are also looked upon.
Step-10: The user gets a message about the present condition and consultations for the future.

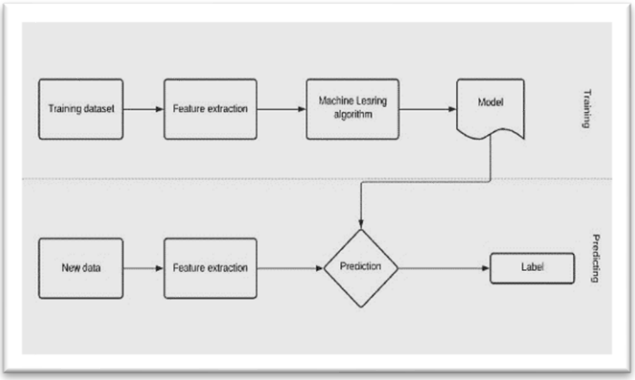


Figure 2(a): Flow diagram of the proposed algorithm.

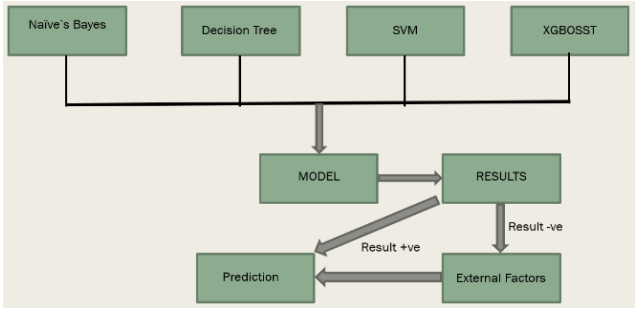


Figure 2(b): Flow diagram of the proposed algorithm with new input arrival

Figure 2(a) explains flow diagram and 2(b) explains the manner of prediction done on each new input arrival. Each new input is taken to individual classifiers namely Naive Bayes, Decision tree, SVM, XGBoost. Then each of the classifier's results, with individual output are predicted which are verified for any ambiguous data or any system error then it is converted into a model.

In the model verification and validation, results are calculated and then the results are categorized into positive and negative. If a patient has a negative result, it redirects to take external factors and then make a prediction according to it, whereas in positive results a warning message is shown and consultation with a specialist is advised.

4 System U/I design

We have designed a webpage using flask for the implementation of the proposed model in Fig 3 (a-d) and its subparts. Here four figures depict the input and result pattern of the heart disease patient.

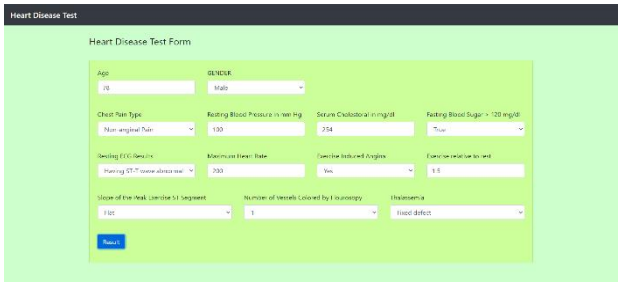


Figure 3(a): Input of heart disease symptoms.

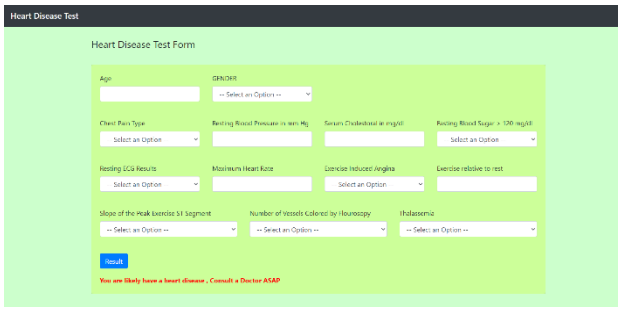


Figure 3(b): Positive result of heart disease symptoms.

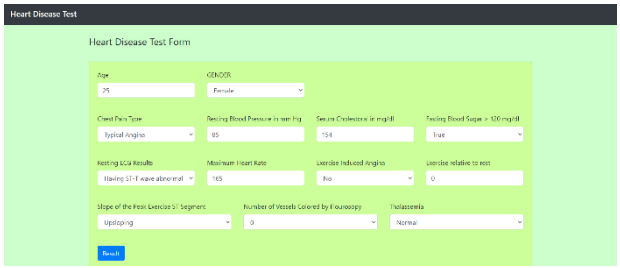


Figure. 3(c): Input of Non-heart disease symptoms.

Figure 3(c): Input of external factors of Non-heart disease symptoms.

Figure 3(d): Result of negative heart disease but having external factors positive.

5 Result and discussion

Table 5 clearly shows that all 4 methods used have resulted very accurately in training and testing areas. The training and testing are done on the ratio of 60 and 40 on the same dataset. Some data are kept reserved for model evaluation and application testing at a later stage to evaluate a proper idea of the system errors. While testing at the latest stage, no error was found either at the system end or at the web end. The resultant accuracy was calculated using a table-6 with the proposed algorithm (Modified hybrid classifier) of the system proposed is 98.73% which is comparatively far better in the context of previous research.

Table 5: Training and testing results.

Classifiers	**Training accuracy (%)**	**Testing accuracy (%)**
Naive Bayes	85	78
Decision tree	100	96
Support vector machine	100	84
eXtreme Gradient Boosting	100	97
Modified hybrid classifier(Proposed)	100	98.73

Table 6: Confusion matrix of modified hybrid classifier.

	Positive	**Negative**
Positive	760	8
Negative	7	300

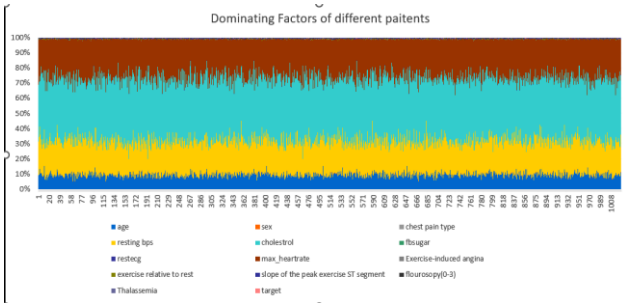


Figure 4: Representation of performance parameter.

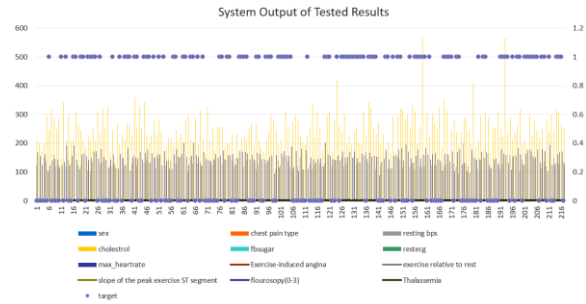


Figure 5: Output of tested results for various parameters.

This graph (shown in Figure 5) refers to the various features entered in the line graph form and the blue dot represents patients' results, the blue dot at 0 means heart diseases not found blue dot at 1 means heart diseases detected. (Note that this is the sample testing done on 216 data for a better understanding of system results and to get an overall view).

The graph (Fig-6) shows us various algorithms which are present in the industry and their accuracy against the proposed method. The Blue dotted lines suggest the industry standards line of accuracy.

The purpose of this experiment was to analyze and predict the possibility of heart disease with high precision which benefits directly to human society. Results shown in the process of prediction suggest high accuracy and fewer system failures. The on-ground implementation of the project has been successfully deployed with accurate precision. At no point in time, no conclusive system error has occurred neither at the system end or at the web application end.

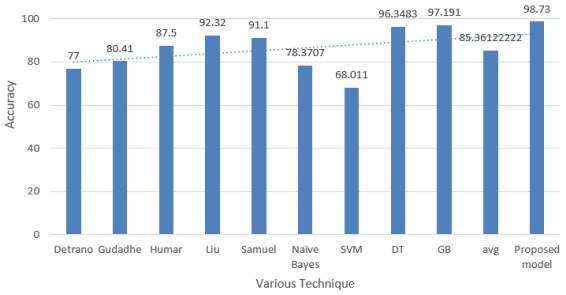


Figure 6: Comparative analysis of algorithms.

6 Conclusion and future scope

Our proposed system achieved an accuracy of 98.73% where the model accepts 13 clinical data and 5 environmental data, and it is trained using backpropagation algorithms to read it and analyze the presence or absence of heart disease in a patient. We also presented a user-friendly web application which helps a patient easily access his/her present condition and act accordingly.

Integrated multiple disease prediction-based models could be designed so that a user can analyze any condition according to their choice. A market review could also be done in order to launch the prototype for medical and general public use. All these may help society to come closer in the fight against modern-day diseases and their detection.

Acknowledgment

We would thank Dr. Mohit Chowdhary (MBBS from Kolkata Medical College, Junior resident, Department of Medicine, All India Institute of Medical Sciences, New Delhi) for the Technical Analysis of the Diseases and for giving the medical point of view of the paper.

Author's Contributions

All the authors have contributed equally to this paper.

References

[1] Medline plus: heart diseases, 2021. http://www.nlm.nih.gov/medlineplus/heartdiseases.html (Accessed on April 22, 2021)

[2] Mohamed, MohamedM G., Mohammed Osman, Babikir Kheiri, Maryam Saleem, Alexandre Lacasse, and Mohamad Alkhouli. "Polypill for cardiovascular disease prevention: systematic review and meta-analysis of randomized controlled trials." International Journal of Cardiology (2022). https://doi.org/10.1016/j.ijcard.2022.04.085

[3] Tsanas, Athanasios, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity." Journal of the royal society interface 8, no. 59 (2011): 842-855. https://doi.org/10.1098/rsif.2010.0456

[4] Detrano, Robert, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. "International application of a new probability algorithm for the diagnosis of coronary artery disease." The American journal of cardiology 64, no. 5 (1989): 304-310. https://doi.org/10.1016/0002-9149(89)90524-9

[5] Gudadhe, Mrudula, Kapil Wankhade, and Snehlata Dongre. "Decision support system for heart disease based on support vector machine and artificial neural network." In 2010 International Conference on

Computer and Communication Technology (ICCCT), pp. 741-745. IEEE, 2010. https://doi.org/10.1109/ICCCT.2010.5640377

[6] Kahramanli, Humar, and Novruz Allahverdi. "Design of a hybrid system for the diabetes and heart diseases." Expert systems with applications 35, no. 1-2 (2008): 82-89. https://doi.org/10.1016/j.eswa.2007.06.004

[7] Li, Yanping, Tianyi Huang, Yan Zheng, Tauland Muka, Jenna Troup, and Frank B. Hu. "Folic acid supplementation and the risk of cardiovascular diseases: a meta-analysis of randomized controlled trials." Journal of the American Heart Association 5, no. 8 (2016): e003768. https://doi.org/10.1161/JAHA.116.003768

[8] McKinley, DeAngelo, Pamela Moye-Dickerson, Shondria Davis, and Ayman Akil. "Impact of a pharmacist-led intervention on 30-day readmission and assessment of factors predictive of readmission in African American men with heart failure." American journal of men's health 13, no. 1 (2019): 1557988318814295. https://doi.org/10.1177/1557988318814295

[9] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." In 2008 IEEE/ACS international conference on computer systems and applications, pp. 108-115. IEEE, 2008. https://doi.org/10.1109/AICCSA.2008.4493524

[10] Olaniyi, Ebenezer Obaloluwa, Oyebade Kayode Oyedotun, and Khashman Adnan. "Heart diseases diagnosis using neural networks arbitration." International Journal of Intelligent Systems and Applications 7, no. 12 (2015): 72. DOI: 10.5815/ijisa.2015.12.08

[11] Liu, Peter Y., Alison K. Death, and David J. Handelsman. "Androgens and cardiovascular disease." Endocrine reviews 24, no. 3 (2003): 313-340. https://doi.org/10.1210/er.2003-0005

[12] Samuel, Oluwarotimi Williams, Grace Mojisola Asogbon, Arun Kumar Sangaiah, Peng Fang, and Guanglin Li. "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction." Expert Systems with Applications 68 (2017): 163-172. https://doi.org/10.1016/j.eswa.2016.10.020

[13] Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." In 2020 international conference on electrical and electronics engineering (ICE3), pp. 452-457. IEEE, 2020. https://doi.org/10.1109/ICE348803.2020.9122958

[14] 14. Zhou, Joey Tianyi, Hao Zhang, Di Jin, Xi Peng, Yang Xiao, and Zhiguo Cao. "Roseq: Robust sequence labeling." IEEE transactions on neural networks and learning systems 31, no. 7 (2019): 2304-2314. https://doi.org/10.1109/TNNLS.2019.2911236

[15] Geweid, Gamal GN, and Mahmoud A. Abdallah. "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique." IEEE Access 7 (2019): 149595-149611. https://doi.org/10.1109/ACCESS.2019.2945527

[16] Ahammad, Tanvir, Tamanna Yesmin, Md Mahmudul Hasan, Sudipta Kumar Mondal, and Selina Sharmin. "An approach for collaboration between different stakeholders to strengthen the public health system." Informatica 46, no. 7 (2022). https://doi.org/10.31449/inf.v46i7.3986

[17] Gadiparthi, Manjunath, and E. Srinivasa Reddy. "Optimizing the Quality of Predicting the ill effects of Intensive Human Exposure to Social Networks using Ensemble Method." Informatica 46, no. 7 (2022). https://doi.org/10.31449/inf.v46i7.4212

[18] Milioris, Konstantinos, Charalampos Konstantopoulos, and Konstantinos Papageorgiou. "Perceptions and needs of health professionals concerning health information systems." Informatica 46, no. 7 (2022). https://doi.org/10.31449/inf.v46i7.3974

[19] Almola, Sahera Abued Sead, Mohammed H. Haloob Alabiech, and Esraa Jasem Harfash. "Citrus Diseases Recognition by Using CNN." Informatica 46, no. 7 (2022). https://doi.org/10.31449/inf.v46i7.4284

Sentiment Analysis and Machine Learning Classification of COVID-19 Vaccine Tweets: Vaccination in the Shadow of Fear-trust Dilemma

Samet Tüzemen[1], Özge Barış-Tüzemen[2] and Ali Kemal Çelik*[1]
1 Department of Business Administration, Ardahan University Çamlıçatak Ardahan, Türkiye
2 Department of Econometrics, Karadeniz Technical University Kalkınma Trabzon, Türkiye
E-mail: samettuzemen@ardahan.edu.tr, ozgebariss@gmail.com, alikemalcelik@ardahan.edu.tr
* Corresponding author

Keywords: COVID-19 vaccine, sentiment analysis, machine learning, text mining, twitter

Received: March 4, 2022

In addition to infecting millions of people and causing hundreds of thousands of deaths, COVID-19 has also caused psychological and economic devastation. Studies on the vaccine, which is considered to be the only way to eliminate this pandemic, have been rapidly completed and more than 10 vaccines have begun to be applied worldwide by 2021. One of the biggest obstacles to the fight against COVID-19 is the hesitation against the vaccine. The fear factor, fed by incomplete and false information spreading rapidly through social media applications such as Twitter, is thought to be the main reason for this hesitation. In this study, the general sentiment against the COVID-19 vaccine is analyzed. For this, in the first week of January 2021, more than 8000 tweets are extracted with R statistical software and Twitter API, and appropriate sentiment analysis methods are applied. On the other hand, accuracy values are obtained by applying Logistic Regression and Naïve Bayes methods, which are effective and widely used supervised machine learning methods, for sentiment classification. Although the results indicate that there is a positive attitude about the vaccine, it is remarkable that the rate of negative sentiments is relatively high (30%). Trust is the dominant sentiment on the positive side, while fear is the dominant sentiment on the negative side. According to the results of the classification methods, accuracy values are close to 90%.

Povzetek: Študija obravnava splošno razpoloženje glede cepiva za COVID-19 na Twitterju.

1 Introduction

COVID-19 is a disease caused by the virus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is transmitted from person to person, affecting the respiratory tract. This disease, which emerged with the detection of the first case in Wuhan province of China in December 2019 and spread all over the world in a few months, was declared a pandemic by the World Health Organization (WHO) on March 11th, 2020 and has been the most important agenda item in the world until today. COVID-19 is transmitted by inhaling droplets emitted by sick individuals during a speech or sneezing/coughing. For this reason, it is recommended to pay attention to social distance, use of masks, and cleanliness as a method of protection from disease. Since these recommendations were found to be insufficient to contain the spread of the disease, governments have implemented various advanced measures such as restrictions and closures.

The measures taken by many countries around the world, in the form of travel and gathering bans and lockdowns, have contributed to overcoming the periods in which the spread of the epidemic accelerated, called waves, and ensured the control of the epidemic to a certain extent, as seen in Figure 1. However, the restriction of economic activity and social life in this dimension has created great pressure on individuals and the economies. Especially individuals who are currently trying to cope with the shock effect of a worldwide epidemic have also faced the loss of social interaction. As a result, individuals have started to experience disorders such as stress, anxiety, and depression [1].

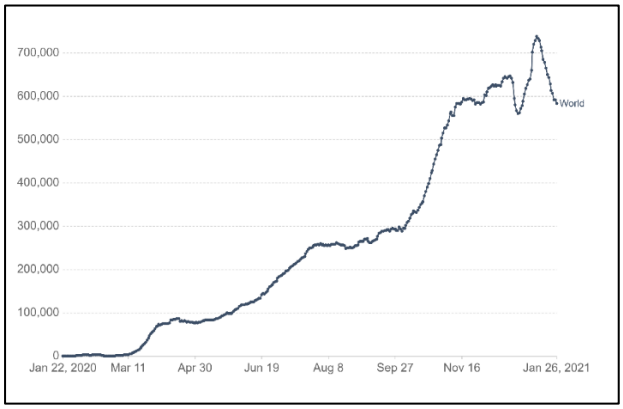


Figure 1: Daily new confirmed COVID-19 cases (world).

On the other hand, these measures had a great impact on macroeconomic indicators such as unemployment and

budget balance. This situation caused a sudden decline in the Gross Domestic Product of the countries, causing almost all of them to experience a large decrease. As seen in Figure 2, the shrinkage experienced in GDP has found an average of 4.4%. In addition, there also have been dramatic increases in unemployment rates. All these negativities experienced in economic terms ultimately caused the psychological conditions of individuals to get even worsen [3].

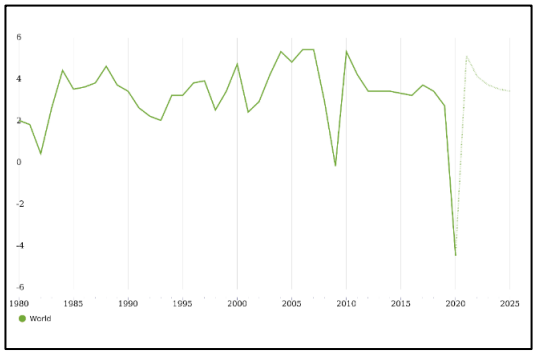


Figure 2: Real GDP growth (Annual %).

Simultaneously with the emergence of the disease, scientists in many parts of the world began working to develop a vaccine that would be effective against the virus. It is known that years, not months, are needed for an effective and safe vaccine to emerge after all procedures are completed. Despite this, a great effort has been made for an effective vaccine that will end the COVID-19 pandemic, and the development process of at least 3 vaccines has been completed before the end of 2020. In some countries such as China, United Kingdom, and Russia, it has even been granted permission to use these vaccines in emergencies. As of January 2021, 10 vaccines have been used by various countries and over 70 million people have been vaccinated. Figure 3 shows the course of vaccination in 10 countries where the most doses are applied.

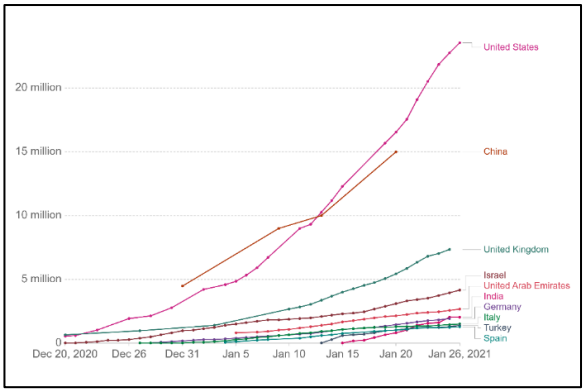


Figure 3: Cumulative COVID-19 vaccines doses administered (highest 10 countries).

Considering the psychological and economic destruction of this pandemic, it is expected that the beginning of the vaccination process, which is likely to end the epidemic, would have a positive effect on people. This virus has infected 100 million people worldwide and caused the death of 2 million people as of January 2021. Despite this phenomenon, the positive attitude towards the vaccine is not high as is expected. Although it varies based on countries, it is observed that there is a remarkable rate of skepticism in the society against the vaccine [5]. The most important factor that triggers this attitude, which is an important obstacle in the effective fight against a pandemic, is the rapid spread of misleading information based on conspiracy theories. Social media has become the primary communication tool that enables information to spread rapidly around the world. However, the accuracy of the aforementioned information cannot always be guaranteed, and this causes information corruption. This situation makes it difficult to manage the perception of society in such an important period. As a result, even the vaccine, which is the world's only hope to end this global crisis, was faced with a significant negative response.

This study aims to reveal the public sentiment against the COVID-19 vaccine as of the first week of 2021 by examining the posts (tweets) from Twitter, which is an important social media tool with a large user base, while the ongoing vaccination activities are given. For this purpose, using the R statistical software, Twitter posts are compiled, and sentiment analysis is performed with the data cleaned with appropriate methods. Then, the efficiency of the established models is examined by applying Logistic Regression and Naïve Bayes Classification methods, which are the most frequently used machine learning methods.

2 Background and literature review

Studies on COVID-19 disease have increased dramatically after spread around the world and declared as a global pandemic by WHO. Not only the medical effects of COVID-19 on sick people, but also the psychological and behavioral effects on the whole society, and even the socio-economic effects on the countries are examined in these studies. Considering the scope and impact of the disease, evaluating these studies independently from each other will prevent one to understand the real dimension of each effect. For this reason, some of the wide-ranging researches are referred and their contribution to this study is examined.

As in every large socio-economic incident, the primary impact of the COVID-19 outbreak has been on the psychology of individuals. Especially, the increasing number of cases and deaths and the restrictions imposed by the governments started to create increasing pressure on the individuals in the society. It is thought that determinants such as education, age, gender, and social

status have an important contribution to the extent of this pressure. Accordingly, certain groups are experiencing unemployment and cost of living pressure, while others must cope with concerns such as education and socialization. Psychological problems such as stress, anxiety disorder, and depression accompany this pressure. With the rapid spread of the disease itself, the spread of true and false information about it on social media has increased the extension of individual traumas. Although efforts are made to alleviate social trauma through various methods such as free online group therapies, some researchers argue that the effects of this trauma will extend into the post-pandemic period and only then will its profound effects be understood [1, 6, 7, 8, 9, 10, 11].

Today, people prefer to share their feelings on social media. For this reason, social media applications such as Twitter have become a very large and important data source to measure the feelings of individuals and societies in the face of certain events. In this process, many studies have been conducted using tweets to investigate how people feel about the COVID-19 outbreak. In these studies, which are called sentiment analysis, researchers have applied various machine learning classification methods such as Logistic Regression, Naïve Bayes, Vector Support Machine, and Recurrent Neural Network, which are widely used today. The results vary according to the demoFigureic structure and the measures taken by the governments. However, it is seen that the presence of high polarity in relatively homogeneous groups, the reaction of individuals to an event is generally directly related to their characteristics [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22].

The impact of COVID-19 on the world economy has been devastating. In addition to the restriction of international transportation and trade, because of the lockdown implementations and the dramatic decline in commercial activity, the economies around the world rapidly entered a severe recession. As a natural consequence of this, unemployment rose to record levels and financial difficulties increased significantly. Although developed and rich countries have taken some measures to mitigate the impact of the pandemic on the general economy and individuals, the crisis has deepened in countries that are already experiencing difficulties. Another reason for the psychological disorders that coped with the COVID-19 pandemic is the mentioned unemployment. While most research show that unemployment is an important trigger of stress and depression, it also emphasizes that this situation causes an increase in suicide cases. Again, the researchers concluded that the groups considered as minorities are more vulnerable when considering the factors of race, gender, and age in the face of the above factors [23, 3, 24, 25, 26, 27, 28, 29].

Experts point out the need for herd immunity to end this pandemic and therefore stop the material and moral losses. Accordingly, in order for the epidemic to slow down and disappear, at least 60% of the population must be immunized [30]. This can be done in two ways. First, this proportion of people should get sick. Second, people should be vaccinated at this rate. There have been countries that have tried the first method at the early stage of the pandemic. However, with the realization that the cost of this is too high to be incurred, the second method has become the only hope for the whole world. Especially the shock wave experienced at the beginning of the pandemic provided support for vaccination studies to a large extent.

With the availability of at least a few different vaccines in the last months of 2020, conspiracy theories that spread rapidly on social media emerged. These conspiracy theories have triggered an unsafe environment for the vaccine. In this context, the hesitation against the COVID-19 vaccine is being studied extensively. [31] argue that the current hesitation against vaccination will not disappear in a short time, even with a devastating pandemic such as COVID-19, and this should be tackled at the local level. [32] found that 71.5% of the participants were willing to get the COVID-19 vaccine in their survey in June 2020, with 13,426 people from 19 countries. [33] measured vaccine literacy and attitude against possible COVID-19 vaccine in their survey study for Italy. Again, the results of this study carried out in June 2020, show that there is an 80-90% positive attitude towards the vaccine.

Examining the size of the hesitation against vaccines in China in May 2020, [34] found that 95% of the participants trust the vaccine to be developed in the country and 83% want to get the COVID-19 vaccine when it is ready. [35] conducted a similar study for the USA in May 2020 and as a result, 69% positive response was obtained for the COVID-19 vaccine. On the other hand, [36] revealed in the survey they conducted for the UK in September 2020 that 54% of the participants had a positive approach to the COVID-19 vaccine. [37], who examined the rate of refusal of the COVID-19 vaccine by 5 consecutive survey studies in France between May and October 2020, found that this rate gradually increased. According to the findings of [38], who conducted similar research for Italy, the situation was in line with the results of the previous study. Researchers have stated that there is a decrease in the intention of vaccination between the two stages of the epidemic in Italy and that the proportion of people who intend to be vaccinated is not enough to end the epidemic (pp. 786-787).

Many similar studies have been conducted for many countries such as Finland, Israel, Pakistan, and

Indonesia. According to the findings and the joint opinion of the researchers, the skeptical attitude towards the COVID-19 vaccine is alarming and urgent action should be taken against it [39, 40, 41, 5, 42, 43]. On the other hand, in the survey study conducted for nursing students, who are healthcare professionals of the future, it was revealed that 63% of the participating students intend to get the COVID-19 vaccine [44]. This rate clearly shows that even young people with health education have a skeptical attitude towards vaccination. When the studies are evaluated as a whole, it is seen that the size of the hesitation against the vaccine is at a worrying level. Almost all researchers agree that a proactive method should be followed in solving this problem.

3 Data set and methodology

In this study, public sentiment analysis about the COVID-19 vaccine is examined. The data set used for this purpose is extracted from Twitter with the keywords "coronavirus vaccine" on January 9th, 2021, using the R statistical software and the "rtweet" library. More than 8000 tweets are converted to lowercase letters, freed from repetitions, punctuations, numbers, stop words, URLs, and non-ASCII words, and finally lemmatized in order to make them ready for analysis. Finally, the remaining 7935 tweets have been converted into a term document matrix.

To analyze the general attitude towards COVID-19 vaccines, the sentiment analysis method, which is a frequently used and effective method in big data analytics, is used. Sentiment analysis is defined as the classification of the main idea in a text with the applications of natural language processing and text analytics. Sentiment analysis aims to understand the attitude of the author by detecting the emotion polarization in a text and classifying it as positive, negative, and neutral ([45] pp. 53-54). For this, a dictionary-based emotion score is determined for each word in the text using flexible and open-source programming languages such as R and Python and related packages. Later, this score, determined on the basis of words, is calculated for the whole text. As a result, the text is classified as positive, negative, or neutral [45].

Despite the important advantage of being less complex, the score calculation method is not efficient enough in some cases. A positive sentence with negative score words will be evaluated as negative by this method. On the other hand, the machine learning approach, which automates processes more, is widely used in sentiment classification. In particular, a sentiment classification model is created by training the available data with supervised machine learning methods, and the obtained accuracy values are compared. This comparison is used to determine if the model has been set correctly, or if there is an overfit or underfit issue. Logistic Regression and Naïve Bayes, which are among these supervised machine learning methods, are used in this study.

Logistic regression is a common type of generalized linear models and models the probability of some events occurring as a linear function of a set of predicted values. In other words, the Logistic Regression method tries to estimate the probability of the dependent variable having a certain value instead of estimating the value of the dependent variable. For example, instead of guessing whether a soccer team will beat the round it played, it tries to predict the probability of passing the round. The actual state of the dependent variable is determined by looking at the estimated probability. If the predicted probability is greater than 0.50, the estimate is closer to YES (i.e., to pass the round), otherwise, the failure to pass the round is more probable. Logistic Regression is used only when the dependent variable is a categorical binary (0 or 1, YES or NO, etc.). In this case, these two possibilities are calculated as $P(y_j = 0) = 1 - p_j$ and $P(y_j = 1) = p_j$with the available data. In this case, the linear logistics (logit) model is established as follows ([46] pp. 157-158).

$$\log\left(\frac{p_j}{[1-p_j]}\right) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_n X_{nj}$$

Naïve Bayes is a simple but effective machine learning classification method that uses the Bayes rule based on the assumption of conditional independence of variables. Bayes theory is a method of calculating the probability of event A to occur depending on event B. It is basically formulated as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where, $P(A)$ and $P(B)$ are the probability of occurrence of events A and B, respectively, $P(B \mid A)$ is the conditional probability of event B to event A, and lastly, $P(A \mid B)$ is the conditional probability of event A to event B. Based on this, the Naïve Bayes classification equation is simply shown as:

$$P(y \mid x_1, \ldots, x_j) = \frac{P(x_1, \ldots, x_j \mid y)P(y)}{P(x_1, \ldots, x_j)}$$

where, $P(y \mid x_1, \ldots, x_j)$ is the posterior conditional probability of class (y) to observation values (x_n), $P(x_1, \ldots, x_j \mid y)$ is the conditional probability of observation values to class. Finally, while $P(y)$ is the prior probability of the class, $P(x_1, \ldots, x_j)$ is called the marginal probability ([47] pp. 279-280).

4 Findings

In this part of the study in which the sentiment analysis about the COVID-19 vaccine is examined, the obtained findings are presented. Accordingly, the frequency distribution of 200 or more words in extracted and cleaned tweets with appropriate methods is presented in Figure 4.

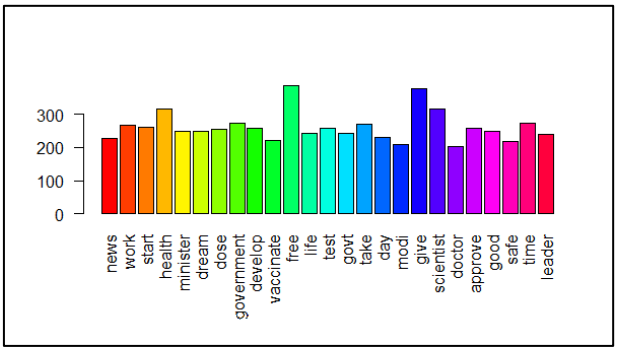


Figure 4: Word frequency of COVID-19 vaccine tweets.

As seen in Figure 4, 4 words namely, free, give, scientist, and health are used more than 300 times. Another important result that is seen from the Figure is that there are words with positive meanings such as good, safe, and approve among the words used more than 200 times, and the absence of words with negative meanings. On the other hand, all the words in the tweets used in the study are presented in Figure 5 in the form of a word cloud.

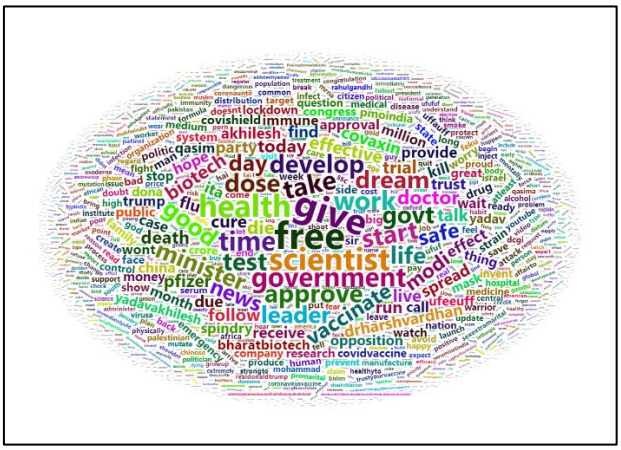


Figure 5: Word cloud of COVID-19 vaccine tweets.

Although it has a lower frequency, there are also negative words such as stop, kill, opposition, and death in tweets, as seen in Figure 5. When evaluated as a whole, the distribution of sentiments in the tweets of this study is shown in Figure 6.

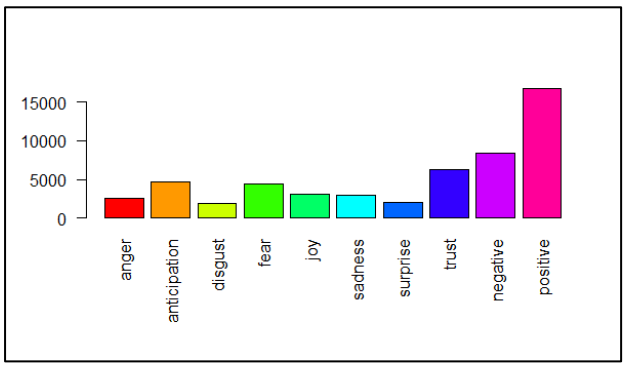


Figure 6: Sentiment frequency distribution of COVID-19 vaccine tweets.

As seen in Figure 6, the strongest sentiment against the COVID-19 vaccine as of the first week of January 2021 is positive. Accordingly, positive sentiments are about twice as much as negative sentiments. With a simple approach, it can be said that approximately 66% of the tweets about the COVID-19 vaccine are positive and approximately 33% are negative. These results coincide with the findings of [35], [36], and [44] examined in the literature section of the study. Although these rates give the impression that there is a positive approach to the vaccine at first glance, it is critically close to 60%, which is required for the immunity rate also known as herd immunity, to end the pandemic. Therefore, it is not wrong to comment that the rate of negative sentiments towards the vaccine is high. On the other hand, it is seen that feelings of trust and hope are dominant on the positive side, and fear is dominant on the negative side.

In order to classify tweets with supervised machine learning methods, the tweets with positive emotion score are marked as 1, and the ones with negative scores are marked as 0. With this marking, the Logistic Regression model is established and applied to the train and test data set separated as 80%-20%. The obtained confusion matrix and accuracy values are presented in Table 1.

Table 1: Confusion matrix and accuracy values for logistic regression.

Train data			**Test data**		
	Actual			*Actual*	
Prediction	**Negative**	**Positive**	**Prediction**	**Negative**	**Positive**
Negative	509	161	**Negative**	129	38
Positive	388	4109	**Positive**	83	1009
Accuracy: ***0,8937***			*Accuracy:* ***0,9039***		

As seen in Table 1, the Logistic Regression model has made the negative and positive classification of the sentiment of tweets with very high accuracy. In addition, the sensitivity value giving the true positive rate is calculated as 0.9637 for the train data set, while the specificity value giving the true negative rate is

calculated as 0.6085. On the other hand, the results of the Naïve Bayes model created for emotion classification of tweets about the COVID-19 vaccine are presented in Table 2.

Table 2: Confusion matrix and accuracy values for naïve bayes.

Train data			**Test data**		
	Actual			*Actual*	
Prediction	**Negative**	**Positive**	**Prediction**	**Negative**	**Positive**
Negative	544	254	**Negative**	132	65
Positive	353	4016	**Positive**	80	982
Accuracy: ***0,8825***			*Accuracy:* ***0,8848***		

As seen in Table 2, where the results of the Naïve Bayes classification model are presented, the accuracy values are close to the results of Logistic Regression. The sensitivity value for the train data set is 0.9405 and the specificity value is 0.6065. When the findings are compared with the results obtained from the Logistic Regression model, it is seen that the Logistic Regression classification model is slightly more effective.

5 Conclusion

The aim of this study is to measure and evaluate the attitude towards the COVID-19 vaccine with social media, which has become the most important communication tool today. For this purpose, the tweets about the vaccine are extracted and sentiment analysis about the vaccine is made with various classification methods. For this, on January 9th, 2021, more than 8000 tweets belonging to the previous week are extracted via R statistical software and Twitter API, and the obtained data set is cleaned through appropriate libraries and made ready for analysis. The results of the sentiment analysis and machine learning classification are shared in the findings section of this study.

When the results of the study are evaluated, it is seen that positive sentiments about the COVID-19 vaccine are more than negative ones. Therefore, the vaccine, which is seen as the best possible solution to the major problems caused by the pandemic, is generally accepted. On the other hand, the high rate of negative sentiments is worrisome. Similar to the study conducted by [48], this rate (more than 30%) is an indication that hesitation against vaccination should be evaluated carefully. The results obtained with the classification of sentiments reveal that the most dominant sentiment among negative sentiments is 'fear'. Thus, in order to ensure that the fight against the COVID-19 pandemic is not interrupted and the desired level of immunity is achieved, those in the public decision-making position must take strategic steps to combat fear and the underlying uncertainty. The most basic way of this is to fight against misinformation that spread rapidly, especially in social media, by sharing effective and accurate information.

Logistic Regression and Naïve Bayes supervised machine learning methods are applied to classify tweets about the COVID-19 vaccine and their effectiveness is determined and compared. According to the findings, both methods have very high classification efficiency. However, positive sentiment classification is more successful than negative sentiment classification in both methods. It can be thought that the reason for this is the way the negative feelings are expressed (using "not good" instead of "bad"). In this study, which tries to take a snapshot of the attitude towards the vaccine in terms of its results, it is recommended to examine the sentiment locally in future studies on this subject and to investigate how certain practices or developments affect the attitude towards the vaccine instantly.

References

[1] Pillay, A. L., & Barnes, B. R. Psychology and COVID-19: impacts, themes and way forward. *South African Journal of Psychology*, 50(2), 148–153, 2020. https://doi.org/10.1177/0081246320937684

[2] Ourworldindata.org. Retrieved from https://ourworldindata.org/covid-cases?country=~OWID_WRL (Accessed: 27.01.2021)

[4] IMF World Economic Outlook (October 2020). Retrieved from https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/WEOWORLD (Accessed: 21.01.2021)

[3] Achdut, N., & Refaeli, T. Unemployment and Psychological Distress among Young People during the COVID-19 Pandemic: Psychological Resources and Risk Factors. *International Journal of Environmental Research and Public Health*, 17(19), 7163, 2020. http://dx.doi.org/10.3390/ijerph17197163

[5] Sallam, M. COVID-19 vaccine hesitancy worldwide: a systematic review of vaccine acceptance rates. *medRxiv*, 2021. https://doi.org/10.1101/2020.12.28.20248950

[6] Marmarosh, C. L., Forsyth, D. R., Strauss, B., & Burlingame, G. M. The psychology of the COVID-19 pandemic: A group-level perspective. *Group Dynamics: Theory, Research, and Practice*, 24(3), 122-138, 2020. http://dx.doi.org/10.1037/gdn0000142

[7] Atalan, A. Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology,

environment and economy-perspective. *Annals of Medicine and Surgery*, 56, 38-42, 2020. https://doi.org/10.1016/j.amsu.2020.06.010

[8] Akat, M., & Karatas, K. Psychological Effects of COVID-19 Pandemic on Society and Its Reflections on Education. *Turkish Studies*, 15, 1-13, 2020. https://doi.org/10.7827/TurkishStudies.44336

[9] Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. *International journal of environmental research and public health*, 17(6), 2032, 2020. https://doi.org/10.3390/ijerph17062032

[10] Aiello, L. M., Quercia, D., Zhou, K., Constantinides, M., Scepanovic, S., & Joglekar, S. How Epidemic Psychology Works on social media: Evolution of responses to the COVID-19 pandemic. *ArXiv*, abs/2007.13169, 2020.

[11] Susič, D., Tomšič, J., & Gams, M. Ranking Effectiveness of Non-Pharmaceutical Interventions Against COVID-19: A Review. *Informatica*, 46(4), 2022. https://doi.org/10.31449/inf.v46i4.4181

[12] Tiwari, P., Pandey, H. M., Khamparia, A., & Kumar, S. Twitter-based Opinion Mining for Flight Service Utilizing Machine Learning. *Informatica*, 43(3), 381-386, 2019. https://doi.org/10.31449/inf.v43i3.2615

[13] Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, 97, Part A, 2020. https://doi.org/10.1016/j.asoc.2020.106754

[14] de Las Heras-Pedrosa, C., Sánchez-Núñez, P., & Peláez, J. I. Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems. *International journal of environmental research and public health*, 17(15), 5542, 2020. https://doi.org/10.3390/ijerph17155542

[15] Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. *IEEE Access*, 8, 181074-181090, 2020. https://doi.org/10.1109/ACCESS.2020.3027350

[16] Barkur, G., Vibha, & Kamath, G. B. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian journal of psychiatry*, 51, 102089, 2020. https://doi.org/10.1016/j.ajp.2020.102089

[17] Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, 11(6), 314, 2020. http://dx.doi.org/10.3390/info11060314

[18] Nemes L., & Kiss, A. Social media sentiment analysis based on COVID-19, *Journal of Information and Telecommunication*, 2020. http://dx.doi.org/10.1080/24751839.2020.1790793

[19] Zhou, J., Yang, S., Xiao, C., & Chen, F. Examination of community sentiment dynamics due to covid-19 pandemic: a case study from Australia. *ArXiv*, abs/2006.12185, 2020.

[20] Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *Journal of medical Internet research*, 22(4), e19016, 2020. https://doi.org/10.2196/19016

[21] Samant, S. S., Murthy, N. B., & Malapati, A. Categorization of event clusters from twitter using term weighting schemes. *Informatica*, 45(3), 2021. https://doi.org/10.31449/inf.v45i3.3063

[22] Chawla, S., & Mehrotra, M. Impact of emotions in social media content diffusion. *Informatica*, 45(6), 2021. https://doi.org/10.31449/inf.v45i6.3575

[23] OECD (2020). Economic Outlook No 108. Retrieved from https://stats.oecd.org/Index.aspx?DataSetCode=EO# (Accessed: 17.01.2021)

[24] Kong, E., & Prinz, D. Disentangling policy effects using proxy data: Which shutdown policies affected unemployment during the COVID-19 pandemic? *Journal of Public Economics*, 189, 104257, 2020. https://doi.org/10.1016/j.jpubeco.2020.104257

[25] Kawohl, W., & Nordt, C. COVID-19, unemployment, and suicide. *The Lancet Psychiatry*, 7(5), 389-390, 2020. https://doi.org/10.1016/S2215-0366(20)30141-3

[26] Bauer, A., & Weber, E. COVID-19: how much unemployment was caused by the shutdown in Germany? *Applied Economics Letters*, 2020. https://doi.org/10.1080/13504851.2020.1789544

[27] Fairlie, R. W., Couch, K., & Xu, H. The Impacts of COVID-19 on Minority Unemployment: First Evidence from April 2020 CPS Microdata. *National Bureau of Economic Research Working*

Paper Series, 27246, 2020. https://doi.org/10.3386/w27246

[28] Raifman, J., Bor, J., & Venkataramani, A. Unemployment insurance and food insecurity among people who lost employment in the wake of COVID-19. *medRxiv:the preprint server for health sciences*, 2020.07.28.20163618, 2020. https://doi.org/10.1101/2020.07.28.20163618

[29] Blustein, D. L., Duffy, R., Ferreira, J. A., Cohen-Scali, V., Cinamon, R. G., & Allan, B. A. Unemployment in the time of COVID-19: A research agenda. *Journal of Vocational Behavior*, 119, 103436, 2020. https://doi.org/10.1016/j.jvb.2020.103436

[30] Randolph, H. E., & Barreiro, L. B. Herd Immunity: Understanding COVID-19, *Immunity*, 52(5), 737-741, 2020. https://doi.org/10.1016/j.immuni.2020.04.012

[31] Dubé, E., & MacDonald, N. E. How can a global pandemic affect vaccine hesitancy? *Expert Review of Vaccines*, 19:10, 899-901, 2020. https://doi.org/10.1080/14760584.2020.1825944

[32] Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., Kimball, S., & El-Mohandes, A. A global survey of potential acceptance of a COVID-19 vaccine. *Nature medicine*, 1–4, 2020. https://doi.org/10.1038/s41591-020-1124-9

[33] Luigi Roberto Biasio, L. R., Bonaccorsi, G., Lorini, C., & Pecorelli, S. Assessing COVID-19 vaccine literacy: a preliminary online survey. *Human Vaccines & Immunotherapeutics*, 2020. https://doi.org/10.1080/21645515.2020.1829315

[34] Lin, Y., Hu, Z., Zhao, Q., Alias, H., Danaee, M., & Wong, L. P. Understanding COVID-19 vaccine demand and hesitancy: A nationwide online survey in China. *PLoS Negl Trop Dis*, 14(12): e0008961, 2020. https://doi.org/10.1371/journal.pntd.0008961

[35] Reiter, P. L., Pennell, M. L., & Katz, M. L. Acceptability of a COVID-19 vaccine among adults in the United States: How many people would get vaccinated? *Vaccine*, 38(42), 6500-6507, 2020. https://doi.org/10.1016/j.vaccine.2020.08.043

[36] Loomba, S., de Figueiredo, A., Piatek, S., de Graaf, K., & Larson, H. J. Measuring the Impact of Exposure to COVID-19 Vaccine Misinformation on Vaccine Intent in the UK and US. *medRxiv*, 2020. https://doi.org/10.1101/2020.10.22.20217513

[37] Hacquin, A. S., Altay, S., de Araujo, E., Chevallier, C., & Mercier, H. Sharp rise in vaccine hesitancy in a large and representative sample of the French population: reasons for vaccine hesitancy. *PsyArXiv*, 2020. https://doi.org/10.31234/osf.io/r8h6z

[38] Palamenghi, L., Barello, S., Boccia, S., & Graffigna, G. Mistrust in biomedical research and vaccine hesitancy: the forefront challenge in the battle against COVID-19 in Italy. *European journal of epidemiology*, 35(8), 785–788, 2020. https://doi.org/10.1007/s10654-020-00675-8

[39] Karlsson, L. C., Soveri, A., Lewandowsky, S., Karlsson, L., Karlsson, H., Nolvi, S., ... & Antfolk, J. Fearing the disease or the vaccine: The case of COVID-19. *Personality and Individual Differences*, 172, 110590, 2021. https://doi.org/10.1016/j.paid.2020.110590

[40] Khan, Y. H., Mallhi, T. H., Alotaibi, N. H., Alzarea, A. I., Alanazi, A. S., Tanveer, N., & Hashmi, F. K. Threat of COVID-19 vaccine hesitancy in Pakistan: the need for measures to neutralize misleading narratives. *The American journal of tropical medicine and hygiene*, 103(2), 603-604, 2020. https://doi.org/10.4269/ajtmh.20-0654

[41] Harapan, H., Wagner, A. L., Yufika, A., Winardi, W., Anwar, S., Gan, A. K., Setiawan, A. M., Rajamoorthy, Y., Sofyan, H., & Mudatsir, M. Acceptance of a COVID-19 Vaccine in Southeast Asia: A Cross-Sectional Study in Indonesia. *Frontiers in public health*, 8, 381, 2020. https://doi.org/10.3389/fpubh.2020.00381

[42] Dror, A. A., Eisenbach, N., Taiber, S., Morozov, N. G., Mizrachi, M., Zigron, A., ... & Sela, E. Vaccine hesitancy: the next challenge in the fight against COVID-19. *European Journal of Epidemiology*, 35(8), 775–779, 2020. https://doi.org/10.1007/s10654-020-00671-y

[43] Al Awaidy, S. T., & Khamis, F. Preparing the Community for a Vaccine Against COVID-19. *Oman medical journal*, 35(6), e193, 2020. https://doi.org/10.5001/omj.2020.130

[44] Kwok, K., Li, K. K., WEI, W., Tang, A., Wong, S., & Lee, S. Influenza vaccine uptake, COVID-19 vaccination intention and vaccine hesitancy among nurses: A survey. *International Journal of Nursing Studies*. 114. 103854, 2021. https://doi.org/10.1016/j.ijnurstu.2020.103854

[45] Luo T., Chen S., Xu G., & Zhou J. Sentiment Analysis. In: Trust-based Collective View Prediction. Springer, New York, NY, 2013. https://doi.org/10.1007/978-1-4614-7202-5_4

[46] Kantardzic, M. *Data mining: Concepts, models, methods, and algorithms (3rd ed.)*. Hoboken, NJ: Wiley-IEEE Press, 2019.

[47] Albon, C. *Machine learning with Python cookbook: Practical solutions from preprocessing to deep learning*. Sebastopol, CA: O'Reilly Media, 2018.

[48] Callaghan, T., Moghtaderi, A., Lueck, J. A., Hotez, P., Strych, U., Dor, A., Fowler, E. F., & Motta, M. Correlates and disparities of intention to vaccinate against COVID-19. *Social science & medicine (1982)*, 113638, 2021. https://doi.org/10.1016/j.socscimed.2020.113638

https://doi.org/10.31449/inf.v47i1.4497 Informatica **47** (2023) 83–96 **83**

Learning the Structure of Bayesian Networks from Incomplete Data Using a Mixture Model

Issam Salman[1], Jiří Vomlel[2]
[1]Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 120 01, Prague, CZ
[2]Institute of Information Theory and Automation of the CA, Pod Vodárenskou věží 4, 182 00, Prague, CZ
E-mail: Issam.Salman@fjfi.cvut.cz, vomlel@utia.cas.cz

Keywords: Bayesian network, belief-noisy-OR, structure learning, incomplete data, EM-mixture

Received: November 8, 2022

In this paper, we provide an approach to learning optimal Bayesian network (BN) structures from incomplete data based on the BIC score function using a mixture model to handle missing values. We have compared the proposed approach with other methods. Our experiments have been conducted on different models, some of them Belief Noisy-Or (BNO) ones. We have performed experiments using datasets with values missing completely at random having different missingness rates and data sizes. We have analyzed the significance of differences between the algorithm performance levels using the Wilcoxon test. The new approach typically learns additional edges in the case of Belief Noisy-or models. We have analyzed this issue using the Chi-square test of independence between the variables in the true models; this approach reveals that additional edges can be explained by strong dependence in generated data. An important property of our new method for learning BNs from incomplete data is that it can learn not only optimal general BNs but also specific Belief Noisy-Or models which is using in many applications such as medical application.

Povzetek: Razvita je metoda za določitev optimalne Bayesove mreže ob nepopolnih podatkih.

1 Introduction

Bayesian networks (BNs) have been used in a variety of applications. The challenge of learning a BN can be categorized into two parts: (1) structural learning, which involves identifying the topology of the BN; and (2) parametric learning, which involves estimating the conditional probabilities for a given network. The challenge of learning the structure of a BN is by far more difficult than the other one. Most methods, such as [1] and [2], require complete data, while in practical applications we are often confronted with values missing from the dataset; this problem regards both parts (1 and 2) mentioned above and affects the performance of the model learning. A record with a missing value should be omitted from the dataset.

An earlier work [3] studied the impact of learning the parameters and the structure of a BN using hard EM and soft EM with a comprehensive simulation study covering incomplete data.

In this paper, we study the problem of learning the optimal BN structure from incomplete data, adopting a new approach of using the product distribution mixture models to handle missing values; the latter will be used with [2] to estimate the missing values and learn the optimal structure. In addition, we show in this paper that our new approach is able to learn the structure of a Belief Noisy-OR (BNO) [4] model from incomplete data.

2 Bayesian network

A Bayesian network encodes a joint probability distribution over a set of random variables $U = \{X_1, X_2, \ldots, X_m\}$. We consider only discrete variables in this work, which is the most common current usage of BNs. A finite set of states of a variable X_i will be denoted by $\mathscr{X}_i$. Conditional probability distributions (CPDs) are attached to each variable in the network. Their purpose is to quantify the strength of the relationships depicted in the BN through its structure: these CPDs mathematically describe the behavior of that variable under every possible value assignment of its parents. Since to specify this behavior one needs a number of parameters exponential in the number of parents, and since this number is typically smaller than the number of variables in the domain, this approach results in exponential savings in space and time.

Formally, a Bayesian network for U is a pair $B = \langle G, \Theta \rangle$. Its first component, G, is a directed acyclic graph whose vertices correspond to the random variables U, and whose edges represent direct dependencies between these variables. The graph G encodes independence assumptions: each variable X_i is independent of its non-descendants given its parents in G. The second component of the pair, namely Θ, represents the set of parameters that quantify the network. It contains parameter $\theta_{x_i|\Pi_{x_i}} = f(x_i|\Pi_{x_i})$ for each possible value x_i of X_i and Π_{x_i} of Π_{X_i}, where Π_{X_i} denotes

the set of parents of X_i in G. Accordingly, a Bayesian network B defines a unique joint probability distribution over U given by:

$$\begin{aligned} F(X_1 = x_1, \ldots, X_m = x_m) &= \prod_{i=1}^{m} F(X_i = x_i | \Pi_{X_i} = \Pi_{x_i}) \\ &= \prod_{i=1}^{m} \theta_{x_i|\Pi_{x_i}} \end{aligned}$$

for each Π_{X_i} which is a parent of X_i.

2.1 Structure learning of BN

Note that a BN can be viewed from two perspectives: as an effective coding of an independence relationship on the one hand, and as an effective encoding of a high-dimensional distribution of probabilities on the other hand.

One option of learning the structure is to rely on the specialists in the field through a conscious and meticulous process of knowledge gathering. This involves training experts in probabilistic graphical modeling, validating expert opinions, and extracting and testing information. This process all too often leads to disagreements among experts and a lack of reliability pertaining to the model. Nonetheless, in many fields, where data is scarce, this is one of the key approaches to model building.

Another mechanism is the automatic derivation of the model based on a data set. It is this machine learning approach (ML) that we follow here (so that we avoid the very rich field of human knowledge acquisition). For a data set $D = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$, where $\mathbf{u}_i$ is an instantiation of all variables in U, the BN structure learning translates to the problem of learning a network structure from D. Suppose $\mathbf{u}$ is complete and discrete. Consequently, finding the optimal Bayesian network is reduced to finding the optimal structure. The optimal structure can be learned by three approaches coming from the area of ML.

The first is the constraint-based approach to structure learning, which takes advantage of the first perspective and attempts to reconstruct a Bayesian network by analyzing data independence. These algorithms require an infinite amount of data to learn independence with certainty; high-order independence tests can be unreliable unless the sample size is truly huge [5]. The second is the score-based approach, which invests in the second perspective and looks for Bayesian networks that adequately describe the available data with the best score. The core of this approach is to assign a score value $s(G)$ to each acyclic directed graph G. The score function defines an overall order (up to equivalences) over the structures in such a way that a structure with a better description of the data is assigned a higher value. The last approach is a hyper-approach, which mixes the two previous approaches together.

2.2 Score-based

Score-based learning is a technique frequently used for determining the optimal structure. In this process, each candidate is assigned a BN score to measure the goodness-of-fit of a structure to the data. The goal of the learning problem is then to find the optimal scoring structure. The score usually measures how well this BN describes the data set D.

Definition (1): Let $B = \langle G, \Theta \rangle$ be a Bayesian network, and let $D = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be a training set, where each $\mathbf{u}_i$ assigns a value to all variables in $\mathbf{U}$. The *MDL* scoring function of a network B given a training data set D, written $MDL(B|D)$, is given by:

$$MDL(G|D) = LL(G|D) - \frac{\log n}{2} |G|$$

where $|G|$ is the number of parameters in the network. The first term represents the loglikelihood, i.e., it measures the model fit. The second term penalizes the model complexity. The penalty term for MDL is greater than that for most other evaluation functions, since optimal networks with the MDL are usually sparser than optimal networks with function scoring. As its name suggests, an optimal network with MDL minimizes the scoring function rather than maximizing it. The Bayesian information criterion (BIC) [6] is a scoring function whose calculation is equivalent to MDL for Bayesian networks, but it is derived on the basis of the models' asymptotic behavior. Where the score is decomposable, it can be written as a sum of the scores of each variable and its parent set:

$$\begin{aligned} BIC(G|D) &= \sum_{i=1}^{m} BIC(X_i|\Pi_{X_i}) \\ &= \sum_{i=1}^{m} \{LL(X_i|\Pi_{X_i}) - Penalty(X_i|\Pi_{X_i})\} \end{aligned}$$

The score-based algorithms' aim is to optimize this score and return the structure G that maximizes it. As the space of all possible structures is at least exponential in the number of variables m, this presents a number of problems. There are $m(m-1)/2$ possible undirected edges and $2^{m(m-1)/2}$ possible structures for every subset of these edges. Moreover, there may be more than one orientation of the edges for each such choice. One popular choice is hill-climbing [7].

2.3 Structural learning with pruning

Statistical testing is a method of reducing the set of potential DAGs. Another approach to reducing this set is to use constraints provided by experts. Besides that, we can use structural constraints similar to in [2]. The structural constraints can be applied locally as long as they include only one node and its parents.

Algorithm 1 represents an approach to learning the optimal structure of a BN using the constraint rules and a decomposable score [8]. The main function of the algorithm is to compute a collection of candidate parent sets for each variable. Then we optimize across this collection by selecting one parent set for each variable, without creating

directed cycles while maximizing the total score. The following theorem can be used to reduce the numbers of the collections for candidate parents.

Lemma 2.1. *Let X_i be a variable and $\Pi^{'}$ be a candidate parent set for X_i. Suppose that $BIC(X_i|\Pi^{'}) < BIC(X_i|\{\})$. Then $\Pi^{'}$ can be safely ignored from the candidate parent sets.*

Proof. The proof uses the decomposability of the BIC score. Let $G^{'}$ and G be DAGs that differ only on the parent set of X_i where $\Pi^{'}$ is the parent set of X_i in $G^{'}$ and Π is the parent set of X_i in G. Suppose $\Pi \subset \Pi^{'}$. Therefore, if $G^{'}$ does not contain directed cycles then G cannot contain them either. This fact, together with $\text{BIC}(X_i|\Pi) > \text{BIC}(X_i|\Pi^{'})$, implies thast $G^{'}$ is not BIC optimal. This statement also holds if the candidate subset is the empty set $\Pi = \{\}$. □

Algorithm 1 Parent sets evaluation for the BN structure learning algorithm

Input:
$\mathscr{D}$: a data set
m: an integer representing the number of variables in $\mathscr{D}$
Output: Accepted sets of parents for each node
Phase 1: initialize the parameters
$g_i = (V,E)$ // A DAG containing a node and its candidate parent set
S_i: BIC score of g_i
Q_i: priority queue of triples (X_i, Π_{X_i}, S_i) ordered by S_i
Phase 2: mscour($X_i, \mathscr{D}$) function to find the min(BIC) score
S_i = the BIC score of g_i where only X_i is included
return S_i
Phase 3: find the accepted Q_i for X_i
Q_i is empty
S^* = mscour($X_i, \mathscr{D}$)
Π_{X_i} is a parent set for X_i
add (X_i, Π_{X_i}, S^*) to Q_i
For each $X_k, k \in \{1,\ldots,m\}$ do:
  add X_k to Π_{X_i}
  S_{ki} = the BIC score of the updated Π_{X_i}
  if($S_{ki} > S^*$)
    add (X_i, Π_{X_i}, S_{ki}) to Q_k
    For each $X_j, j \in \{1,\ldots,m\}, i \neq j \neq k$, do:
      add X_j to Π_{X_i}
      S_{ki} = the BIC score of the updated Π_{X_i}
      if($S_{ki} > S^*$)
        add (X_i, Π_{X_i}, S_{ki}) to Q_k
      else delete X_j from Π_{X_i}
    end for
  else delete X_k from Π_{X_i}
end for

The $g_i = (V,E)$ in Phase 1 from Algorithm 1 is a DAG containing the set of nodes $V = \{X_i, \Pi_{X_i}\}$ and the set of arcs $E = \{(X_o, X_i), \forall X_o \in \Pi_{X_i}\}$. Algorithm 1 considers all possible parent sets that can lead to an optimal BN. Its implementation is based on [8]. After Phase 3, we find a DAG with the highest BIC from among the variables given the candidate parent sets of each variable. That is done using GOBNILP [9] tool[1] which is a smart algorithm using integer linear programming. We will refer to this algorithm as A1.

Let us also note that, if a dataset is generated from a BN having the empty graph as its structure and this dataset is large enough then, for any parent set $\{\Pi_{X_i} \neq \phi\}$, it holds that $BIC(X_i|\{\}) > BIC(X_i|\Pi_{X_i})$. This implies that the variables are independent and the penalty for larger parent sets makes the BIC value worse for all nonempty parent sets.

One of the axioms of the pruning rules stated in the literature states that if a candidate subset has a better score than another candidate set and the first candidate set is a subset of the second candidate set, it is safe to disregard that second candidate set due to the decomposability of score functions. We have applied the pruning rule as formalized in the theorem 2.1 in Algorithm 1. That algorithm will reduce the collection of accepted parent sets for each node by discarding all parent sets which do not meet the criteria.

3 Incomplete datasets

One of the widespread problems in data mining and machine learning is incomplete data. Values may be missing even from training instances. Nowadays more and more datasets are available, but most of them are incomplete. Therefore, machine learning must cope with this problem. Normally, to learn the BN structure using A1 algorithm [2], we need complete data, such that all instances $\mathbf{u}_i \in D,\ i \in \{1,\ldots,n\}$ are complete and don't have any missing values. In the case of incomplete data and an instance which has a missing value, A1 does not use this instance in the BN structure learning.

3.1 Product distribution mixtures to handle incomplete data

Because of incomplete data, most methods in machine learning cannot be applied. An easy way to deal with this problem is completing the data by simply omitting the incomplete vectors or removing the incomplete variables. But this approach has a weakness: we may lose a massive part of the available information. Another alternative is to use an estimation to replace the missing values [10] (i.e., put in estimates of the missing values). However, for certain reasons, the estimated values have to be typical, and the natural variability of the data will be partially restricted. For that, the product mixture model gives us a better way to directly apply the EM algorithm to complete the dataset [11]. We will refer to this approach as EM-Mixture.

Considering finite mixtures we assume that:

[1] https://www.cs.york.ac.uk/aig/sw/gobnilp/

$$P(X) = \sum_{j=1}^{r} w_j F(X|j), \quad (1)$$

$$F(X|j) = \prod_{i=1}^{m} F_i(X_i|j), \quad \sum_{j=1}^{r} w_j = 1 \quad (2)$$

where $w_j > 0$ is a probabilistic weight of the j-th mixture component, F_i is the conditional distribution of the variable $X_i, i \in 1,\ldots,m$, and r is the number of components. Note that the product components do not imply that the involved variables are independent. In this sense, the mixture model (1) is not restrictive [12]. It is easy to verify that, by increasing the number of components r, we can describe any discrete probability distribution in the form (1). In our experiments, it was selected based on the number of variables in a dataset.

To estimate the mixture parameters, we maximize the log-likelihood function:

$$LL = \sum_{k=1}^{n} \log P(\mathbf{u}^{(k)})$$

where n is the number of records in the dataset D and $\mathbf{u}^{(k)}$ is the k-th datavector from D. We will use the EM algorithm to maximize the log-likelihood function.

Next, we explain how the learned product mixture model will be used to fill in the missing values. Let $C = \{i_1, i_2, \ldots, i_k\}$ be a subset of $M = \{1,2,\ldots,m\}$ such that the corresponding sub-vector

$$\mathbf{u}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k})$$

is complete. Then, under the product mixture model, we can compute the marginal probability of $\mathbf{u}_C$ as

$$P_C(\mathbf{u}_C) = \sum_{j=1}^{r} w_j F_C(\mathbf{u}_C|j) \quad (3)$$

$$F_C(\mathbf{u}_C|j) = \prod_{i \in C} F_i(x_i|j) \ . \quad (4)$$

Let z be an index of a variable unobserved in $\mathbf{u}$, i.e., $z \in M \setminus C$. Under the product mixture model, we can compute the conditional distribution of the missing value $\mathbf{u}_z$ given the complete part $\mathbf{u}_C$ with $P_C(\mathbf{u}_C) > 0$ as

$$P_{z|C}(\mathbf{u}_z|\mathbf{u}_C) = \frac{P_{z,C}(\mathbf{u}_z, \mathbf{u}_C)}{P_C(\mathbf{u}_C)} = \sum_{j=1}^{r} W_j(\mathbf{u}_C) F_z(\mathbf{u}_z|j)$$

where $W_j(\mathbf{u}_C)$ are the conditional component weights:

$$W_j(\mathbf{u}_C) = \frac{w_j F_C(\mathbf{u}_C|j)}{\sum_{j=1}^{r} w_j F_C(\mathbf{u}_C|j)} \ .$$

We thus compute the probability distribution $P_{z|C}(\mathbf{u}_z|\mathbf{u}_C)$ for each missing value of each data vector $\mathbf{u} \in D$ with a missing value. There are several ways of using this probability distribution to fill in the missing value of X_z in $\mathbf{u}$ – in this paper, we select value $\mathbf{u}_z$ maximizing $P_{z|C}(\mathbf{u}_z|\mathbf{u}_C)$ over all values of X_z.

The last step of our presentation is the description of adapting the EM algorithm for learning product mixture models such that it is applicable to incomplete data. Given a data vector $\mathbf{u} \in D$ and a variable X_i with index $i \in \{1,2,\ldots,n\}$, let $\mathscr{N}(\mathbf{u})$ be the subset of indices of the available variables (i.e., observed in that data) of $\mathbf{u}$, and $D(i) \subset D$ be the subset of vectors with observed values of variable X_i:

$$\mathscr{N}(\mathbf{u}) = \{v \in \{1,2,\ldots,n\} : \mathbf{u}_v \text{ observed in } \mathbf{u}\}$$
$$D(i) = \{\mathbf{u} \in D : i \in \mathscr{N}(\mathbf{u})\}$$

In Algorithm 2, we present the modification of the EM algorithm for the product mixture model for incomplete data. For $x_v \in \mathscr{X}_v$, $v \in \{1,2,\ldots,n\}$, and $j = 1,\ldots,r$, we use $F_v(x_v|j)$ to denote the conditional probability of observing value x_v of variable X_v given the component j. The initialization of the EM-Mixture algorithm (presented in Algorithm 2) is performed using the partitions obtained from agglomerative hierarchical clustering implemented in the function *hc* of the R package *mclust* [13]. In our algorithm, the symbol $\delta(x,y)$ denotes the standard delta function equal to one if $x = y$ and equal to zero otherwise.

Algorithm 2 EM-Mixture

Input: D is a data set
Output: a completed data set
Phase 1: initializing:
$w_j,\ j = 1,\ldots,r$
$F_v(x_v|j)$, for $x_v \in \mathscr{X}_v$, $v \in \{1,2,\ldots,n\}$, and $j = 1,\ldots,r$
$L = -\infty$
Phase 2:
repeat
E-Step:

$$q(j|\mathbf{u}) = \frac{w_j \prod_{v \in \mathscr{N}(\mathbf{u})} F_v(\mathbf{u}_v|j)}{\sum_{l=1}^{r} w_l \prod_{v \in \mathscr{N}(\mathbf{u})} F_v(\mathbf{u}_v|l)}, \text{ for } \mathbf{u} \in D,\ j = 1,\ldots,r$$

$$w_j = \frac{1}{|D|} \sum_{\mathbf{u} \in D} q(j|\mathbf{u}), \text{ for } j = 1,\ldots,r$$

M-Step: for $x_v \in \mathscr{X}_v$, $v \in \{1,2,\ldots,n\}$, and $j = 1,\ldots,r$

$$F_v(x_v|j) = \frac{\sum_{\mathbf{u} \in D(v)} \delta(x_v, \mathbf{u}_v) \cdot q(j|\mathbf{u})}{\sum_{\mathbf{u} \in D(v)} q(j|\mathbf{u})}$$

$$L' = \sum_{\mathbf{u} \in D} \log \left[\sum_{j=1}^{r} w_j \prod_{v \in \mathscr{N}(\mathbf{u})} F_v(\mathbf{u}_v|j) \right]$$

$$\mathscr{Q} = L' - L$$

$$L = L'$$

until $\mathscr{Q} \leq \varepsilon$

The EM algorithm converges monotonically to a local

or global maximum or a saddle point of the log-likelihood function L in the sense that the sequence of $\{L^t\}_{t=0}^{\infty}$ does not decrease. The presence of a local maximum makes the starting point of the procedure influential; hence it is selected at random. We use the value of $\varepsilon = 0.005$ to terminate the main loop of the algorithm. The sequence of log-likelihood values generated by E-Step and M-Step is non-decreasing [11] (i.e., $L^{'}L$).

We adapt the BN structure learning algorithm A1 so that it can learn from incomplete data. We use the EM-Mixture algorithm, i.e., Algorithm 2, to make the incomplete data complete in Phase 3. The whole algorithm will be referred to as A2.

3.2 Experiments

The experiments have been repeated ten times on ten different subsets in each MCAR rate on different models, using the generated datasets from the true models summarized in Table 8 in A. We have compared our approach denoted as A2 with three other methods. By A1 we denote the BIC optimal learning from complete data created by omitting all rows containing a missing value. In [3], the authors proposed the soft and hard EM algorithms to fill in the missing values and learn an optimal BN structure from the completed data by Tabu search [14], which we refer to as A3 and A4, respectively.

The test scenarios, which include more than 700 incomplete datasets, are summarized in Figure 1. The resulting BNs of the simulations within each scenario are shown in Tables 1, 2, and 3.

The decision tree shown in Figure 1 is intended to guide practitioners as to which imputation algorithm appears to perform the best, depending on the characteristics of their problem with incomplete data. Each leaf of the decision tree corresponds to a subset of the scenario that we studied, grouped according to the values of the experimental factors, to recommend which algorithm has the best average Structure Hamming Distance [15] (SHD) values between the essential graph of the learned model and the essential graph of the true model. The dominance of the algorithms has been tested using the Wilcoxon test [16]. We say that an algorithm is better than another if it has a lower average SHD and their confidence intervals do not overlap, i.e., the p-value of the Wilcoxon test is lower than 5%.

In the results based on the SHD, A2 has scored the best results. For the results based on the SHD and the Wilcoxon test, we have observed some important general trends:

- A2 appears to be a good algorithm in all scenarios.
- A2 is significantly better than other Algorithms for Model M2 in Leaves B and G.
- A2 is significantly better than other Algorithms for the model Child in Leaf C.
- A2 and A3 are significantly better than A1 and A4 for Models M1 in Leaves C, D, P, and K.

Table 1: Recommended algorithm by decision tree leaf where MCAR rate in [5 - 10] -Group 1.

Leaf	Size	Bayesian Network	Recommended Algorithm
A	Size >5000	Weather	A1, **A2**, A3, A4
		M1	**A2**, A3, A4
B	Size in [3000 - 5000]	Weather	**A2**, A3, A4
		M1	**A2**, A3, A4
		M2	**A2**
		Child	**A2**, A3, A4
C	Size in [1500 - 2500]	Weather	**A2**, A3, A4
		M1	**A2**, A3
		M2	**A2**, A3, A4
		Child	**A2**
D	Size <1000	Weather	**A2**, A3, A4
		M1	**A2**, A3
		M2	**A2**
		Child	**A2**, A3

- A1 is significantly worse than other Algorithms in all scenarios where the data size is smaller than 5,000.

Figure 2 represents the algorithm results of all models with all dataset sizes and all MCAR rates.

Table 2: Recommended algorithm by decision tree leaf where MCAR rate in [15 - 25] - Group 2.

Leaf	Size	Bayesian Network	Recommended Algorithm
E	Size >5000	Weather	A1, **A2**, A3, A4
		M1	**A2**, A3, A4
F	Size in [3000 - 5000]	Weather	**A2**, A3, A4
		M1	**A2**, A3, A4
		M2	**A2**, A3
		Child	**A2**, A3
G	Size in [1500 - 2500]	Weather	**A2**, A3, A4
		M1	**A2**, A3, A4
		M2	**A2**
		Child	**A2**, A3
P	Size <1000	Weather	**A2**, A3, A4
		M1	**A2**, A3
		M2	**A2**
		Child	**A2**, A3

Table 3: Recommended algorithm by decision tree leaf where MCAR rate in [35 - 50] - Group 3.

Leaf	Size	Bayesian Network	Recommended Algorithm
H	Size >5000	Weather	A1, **A2**, A3, A4
		M1	**A2**, A3, A4
G	Size in [3000 - 5000]	Weather	**A2**, A3, A4
		M1	**A2**, A3
K	Size in [1500 - 2500]	Weather	**A2**, A3, A4
		M1	**A2**, A3
M	Size <1000	Weather	**A2**, A3, A4
		M1	**A2**

4 Belief Noisy-Or model

The Belief Noisy-Or (BNO) model is suitable for describing a specific class of uncertain relationships in Bayesian networks [4] common in several practical applications of BNs. As an example, let us mention the QMR-DT network [17]. In Figure 3 we present the structure of a CPT

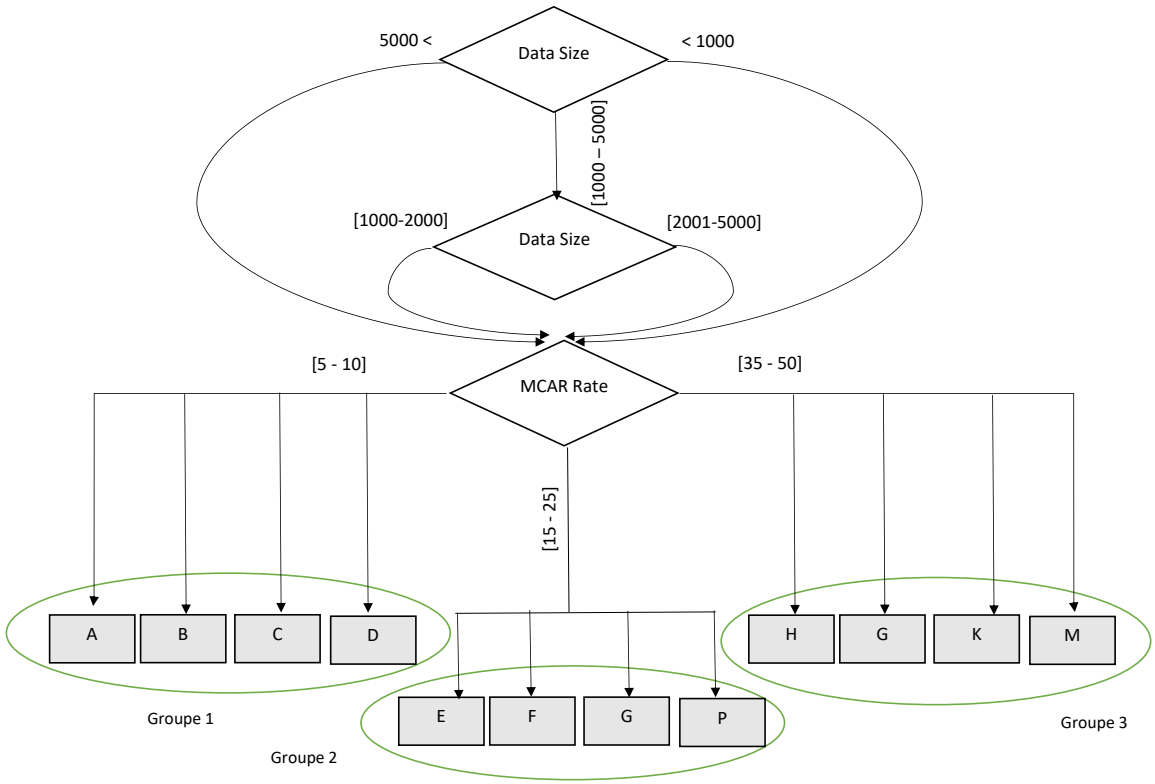


Figure 1: The decision tree for different test scenarios.

$F(Y|X_1,\ldots,X_n)$ where auxiliary nodes $X'_1,\ldots,X'_n$ are added to explicitly separate the noisy relations from the logical OR relation. For a CPT with multiple parent variables $X_1,\ldots,X_n$ the noisy-or is defined as follows[2]:

$$\begin{aligned} F(X_i^{'}=0|X_i=0) &= 1-\alpha, \quad F(X_i^{'}=1|X_i=0)=\alpha \\ F(X_i^{'}=0|X_i=1) &= p_i, \qquad F(X_i^{'}=1|X_i=1)=1-p_i \end{aligned}$$

where $i \in \{1,\ldots,n\}$ and $p_i \in [0,1]$ is the parameter which defines the probability that the positive value x_i of variable X_i is inhibited – it is referred to as the inhibition probability and the parameter α specifies the possibility of a positive value even if the value of the corresponding parent variable is negative. In most experiments, we will set $\alpha = 0$. The CPT of $F(Y|X'_1,\ldots,X'_n)$ represents the deterministic logical OR function, i.e.,

$$F(Y=0|X'_1=x'_1,\ldots,X'_n=x'_n) = \begin{cases} 1 & \text{if } x'_1=0,\ldots \\ & x'_n=0 \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the CPT of $F(Y|X_1,\ldots,X_n)$, which represents the noisy-or function, is computed as follows:

$$\begin{aligned} F(Y=0|X_1=x_1,\ldots,X_n=x_n) &= \prod_{i=1}^{n} F(X_i^{'}=0|X_i=x_i) \\ &= \prod_{1}^{n} (p_i)^{x_i}(1-\alpha)^{1-x_i} \\ F(Y=1|X_1=x_1,\ldots,X_n=x_n) &= 1-\prod_{1}^{n} (p_i)^{x_i}(1-\alpha)^{1-x_i} \end{aligned}$$

4.1 Analysis of BNO models

In this Section, we analyze the BNO models represented in Table 9 in A where $\alpha = 0$. Tables 5, 6, and 7 show the marginal probability distributions (MPD) of the variables in BNO models N1, N2, and BN2O, respectively; look at Figures 11 and 12. The Tables illustrate the decrease of the marginal probability values for F($C_i = 0$) in the case of a node having more than one parent. See Table 7. This decrease is due to the properties of the product of probabilities in (5). On the other hand, they also illustrate the increase of that marginal probability with a higher number of its predecessors in previous layers; that increase depends on the number of layers above and also on the numbers of the edges in those layers. See Table 5 and Table 6.

Using the conditional probability distributions of the variables given their parents, we can easily calculate joint

[2] In the case of one parent variable, we use probability values as specified in Table 4.

Table 4: $F(X_i^{'}|X_i)$ table

		X_i	
		0	1
$X_i^{'}$	0	$1-\alpha$	0.2
	1	α	0.8

Table 5: N1 (Figure 11): Marginal probability distributions

	C1	C2	C3	C4	C5	C6
$F(C_i=0)$	.5	.6	.68	.744	.795	.837
$F(C_i=1)$	.5	.4	.32	.256	.205	.163

Table 6: N2 (Figure 11): Marginal probability distributions

	C1	C2	C3	C4	C5	C6
$F(C_i=0)$	.5	.6	.536	.539	.716	.707
$F(C_i=1)$	.5	.4	.464	.461	.284	.293

Table 7: BN2O (Figure 12): Marginal probability distributions

	C1	C2	C3	C4	C5	C6	C7	C8
$F(C_i=0)$	.5	.5	.5	.5	.5	.129	.36	.36
$F(C_i=1)$	.5	.5	.5	.5	.5	.871	.64	.64

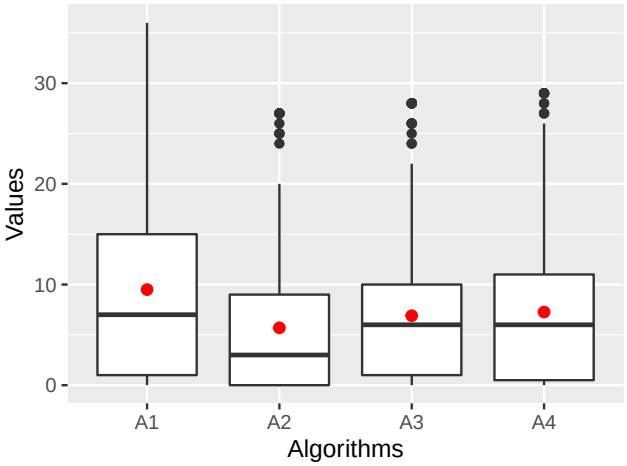


Figure 2: The Structural Hamming Distance to the true models from the resulting models of the structure learning algorithms using data generated from all models, summarized in Table 8 averaged over all data sizes and all MCAR rates.

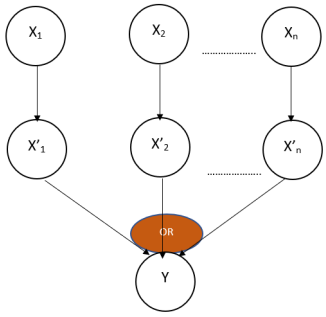


Figure 3: Noisy-or

probability distributions $F(\mathbf{U})$ using formula (1) and conditional probability distributions (CPD) $F(\mathbf{X}_A|\mathbf{X}_B)$, where $\mathbf{X}_A \subseteq \mathbf{U}$ and $\mathbf{X}_B \subseteq \mathbf{U} \setminus \mathbf{X}_A$. Recall that a CPD for a particular configuration $\mathbf{x}_B$ of parent nodes $\mathbf{X}_B$ can be computed as[3]:

$$F(\mathbf{X}_A|\mathbf{X}_B=\mathbf{x}_B) = \frac{F(\mathbf{X}_A,\mathbf{X}_B=\mathbf{x}_B)}{F(\mathbf{X}_B=\mathbf{x}_B)} \quad (5)$$

The Kullback-Leibler Distance (KLD) of two conditional probability distributions $F(\mathbf{X}_A|\mathbf{X}_B)$ and $G(\mathbf{X}_A|\mathbf{X}_B)$ defined on the same state space is computed as the weighted average KLD of $F(\mathbf{X}_A|\mathbf{X}_B=\mathbf{x}_B)$ and $G(\mathbf{X}_A|\mathbf{X}_B=\mathbf{x}_B)$ over all configurations $\mathbf{x}_B$:

$$\begin{aligned} D(F(\mathbf{X}_A|\mathbf{X}_B)||G(\mathbf{X}_A|\mathbf{X}_B)) &= \sum_{\mathbf{x}_B} F(\mathbf{X}_B=\mathbf{x}_B) \\ &* \sum_{\mathbf{x}_A} F(\mathbf{X}_A=\mathbf{x}_A|\mathbf{X}_B=\mathbf{x}_B) \\ &* \log \frac{F(\mathbf{X}_A=\mathbf{x}_A|\mathbf{X}_B=\mathbf{x}_B)}{G(\mathbf{X}_A=\mathbf{x}_A|\mathbf{X}_B=\mathbf{x}_B)} \\ &= \sum_{\mathbf{x}_A,\mathbf{x}_B} F(\mathbf{X}_A=\mathbf{x}_A,\mathbf{X}_B=\mathbf{x}_B) \\ &* \log \frac{F(\mathbf{X}_A=\mathbf{x}_A|\mathbf{X}_B=\mathbf{x}_B)}{G(\mathbf{X}_A=\mathbf{x}_A|\mathbf{X}_B=\mathbf{x}_B)} \end{aligned}$$

We will use KLD of conditional probability distributions estimated from the true data to support our arguments when we explain the results.

4.2 Experiments

We have performed experiments on different Belief noisy-or (BNO) models with their CPTs defined in Table 4 where $\alpha \in \{0, 0.2\}$ and the CPT of a node which has no parent is uniform, i.e., $F(X_i=1|\{\})=0.5, F(X_i=0|\{\})=0.5$. The experiments have been repeated ten times on ten different datasets generated from BNO models with different MCAR rates as specified in Table 9 in Appendix A. In all Figures, we will denote additional edges by blue dashed lines, missing edges by red lines, and edges with different arrows by orange lines.

[3] Please, note that all BNs considered in this paper satisfy the condition $F(\mathbf{X}_B=\mathbf{x}_B)>0$ for all $\mathbf{x}_B$.

4.2.1 Model N1

The true N1 model is shown in Figure 11 in Appendix A. We use this model as an example of a simple model with a chain structure. This model is motivated by some applications, e.g., from telecommunications. Let us summarize the results of this model:

- All algorithms learn the true structure when $\alpha \neq 0$ in all data sizes and all MCAR rates.
- The algorithms A2, A3, and A4 learn structures different from the true model in some cases with $\alpha = 0$, MCAR rate 15% and data size of 1,000. For example, A3 and A4 learn additional edge $C2 \rightarrow C4$, also, A2 learn $C4 \rightarrow C6$ instead of $C5 \rightarrow C6$.
- Using equation (5), we calculate $F(C6|C5)$ and $F(C6|C4)$ from the true model N1. We have found that their KLD value (computed using equation (4.1)) is very small, it is only 0.001. Also, the chi-square test of independence, whose p-value is smaller than 0.0001, reveals that there is a strong dependence between $C6$ and $C4$ in addition to the relationship between $C6$ and $C5$, already explicitly present in the true model. Also, the BIC of the learned structure[4] is -2,252.93 and the BIC of the true model from the same dataset is -2,255.64. This can be explained by the deterministic conditional distribution $F(C6|C5 = 0)$ for $C5 = 0$. For these reasons, we can conclude that we can accept that A2 has learned $C4 \rightarrow C6$ instead of $C5 \rightarrow C6$.

4.2.2 Model N2

The true N2 model is shown on the right hand side of Figure 11 in Appendix A. We use this model as an example of a model more complicated than the previous model N1. This model is motivated by some applications, e.g., by computer networks. We summarize the results of the experiments performed with this model:

- Figure 4 represents the Structure Hamming Distance (SHD) for all tested MCAR rates and models with $\alpha = 0$. We can observe that, as expected, the algorithm's performance is getting better with increasing the data size.
- We can see that A2 on average has a smaller SHD distance to the true model than other algorithms.
- In Figure 5, we compare the models learned from the datasets of size 5,000 with MCAR rate 10% using all four algorithms. We can see that A2 and A3 have the same SHD but they differ in that A3 has a missing edge $C4 \rightarrow C5$ while A2 has an additional edge $C3 \rightarrow C6$. This additional edge can be explained by observing that there is a chain of nodes $C3 \rightarrow C4 \rightarrow C6$ which the state 0 is propagated through because of $\alpha = 0$. In other words, we calculate $F(C3, C6|C1, C2, C4, C5)$ and the product $F(C3|C1, C2, C4, C5) \cdot F(C6|C4, C5)$ from the true model 11 using equation (5). The KLD value (computed as explained in (4.1)) of these two distributions is very small, it is only 0.02. Also, the chi-square test of independence of $C3$ and $C6$ reveals these variables are dependent (the test's p-value is smaller than 0.0001). The additional edge can be also supported by a comparison of BIC values of the learned structure with and without the additional edge $C3 \rightarrow C6$; they are -9,813.67 for the model with the additional edge and -9,880.5 for the true model.
- If $\alpha > 0$ then no additional edge is learned anymore, no matter what the MCAR rate is. Algorithms A2 and A4 we are always able to learn the true structure when the data size exceeds 1,000. Also, A1 and A3 learn the true structure when the data size is larger than 1,500.

[4]We report the BIC value for one of ten datasets since the results for the remaining nine are similar.

4.2.3 BN2O models

These models are motivated by health-care applications, for example by the QMR-DT network [17]. We created 60 different BN2O models consisting of two layers with $N = 20$ nodes in total. They differ in the numbers of nodes in the first layer, namely $L_1 \in \{5, 8, 12, 15\}$; the numbers of nodes in the second layer are $L_2 = 20 - L_1$. The numbers of edges between layers are generated randomly with three different options $\frac{N}{2}$, $\frac{2 \cdot N}{2}$, and $\frac{4 \cdot N}{2}$; each option repeated five times. Using these models, we have generated multiple incomplete datasets where the sizes of the datasets are 3,000 and 5,000 and the MCAR are 10% and 15%. Figure 8 shows the boxplot of additional and missing numbers of edges learned in all instances for each algorithm where the dataset size is 3,000, and for all MCAR rates. The results show that A2 has better results on average (i.e., the distance to the true model is smaller) than other algorithms.

Next we discuss in more detail one simpler example of a BN2O. The structure of this model is shown in Figure 12 in Appendix A.

- Figure 6 represents the SHD of all learned models grouped by MCAR rates with models where $\alpha = 0$.
- The learned models from the dataset of size 5,000, MCAR rate 10% and $\alpha = 0$ using all algorithms are shown in Figure 7. We can see that A2 performs better (i.e., the SHD distance to the true model is smaller) than other algorithms.
- Note the additional edge $C7 \rightarrow C8$ learned by A2 for most datasets. The argument supporting this additional edge is similar to that valid for the additional edge in the N2 model. Again, we can see that KLD of $F(C7, C8|C2, C5)$ and the product $F(C7|C2, C5).F(C8|C2, C5)$ is very small; it is only 0.002. Also, the chi-square test of independence of $C7$

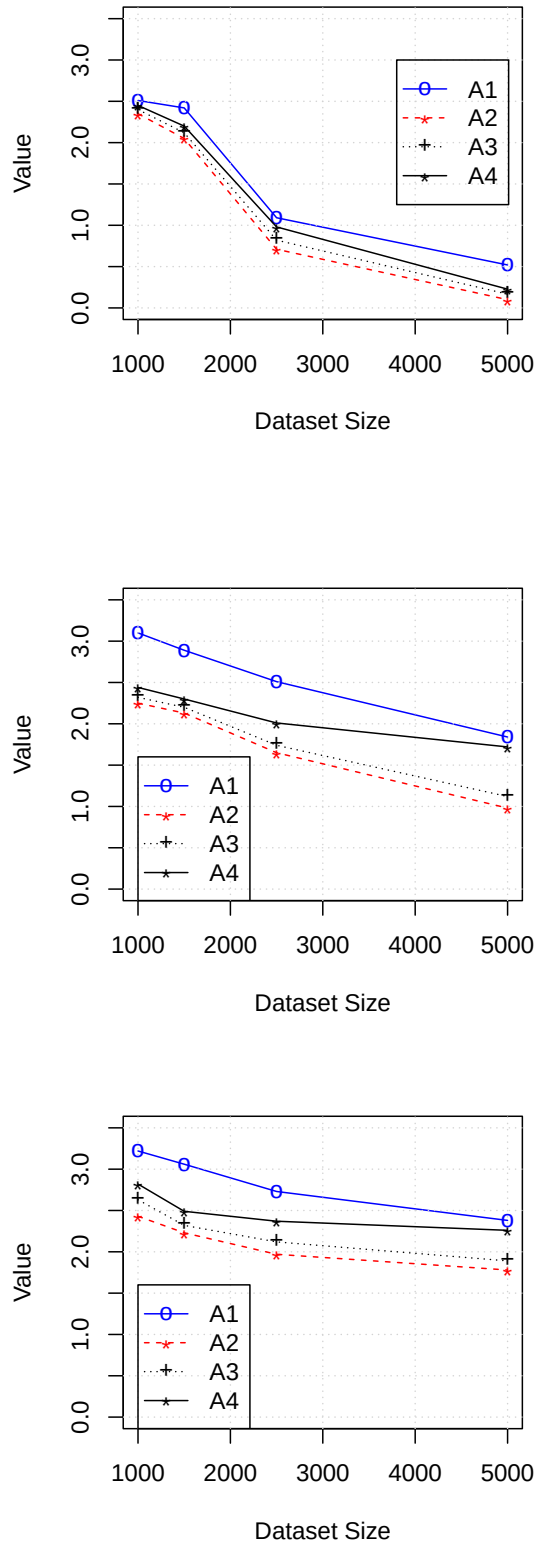


Figure 4: The Structural Hamming Distance of the resulting models of the structure learning algorithms to the true model (with $\alpha = 0$) using the data generated from the N2 model (the true model is presented in Figure 11) using the average over ten experiments for different data sizes and for the MCAR rates of 5%, 10%, and 15%, respectively.

and $C8$ has the p-value smaller than 0.0001 and there is always only a very small difference between BIC of the model with the extra edge and the true model; for example,, BIC of the model with the extra edge is -7,331.8 while the BIC of the true model is -7,338.5 for one of the en generated datasets.

– In the experiments with models having $\alpha > 0$ no additional edge has been learned and the true model is learned successfully when the data size is 2,500 or larger for all MCAR rates.

4.2.4 A large BN2O model

We have performed experiments with a model shown in Figure 13 in Appendix A. This model consists of 25 variables; 14 in the first layer and 11 in the second layer. All algorithms required a data size of more than 3,000 to give a good performance. With the data size of 3,000 (and the MCAR rate of 10%) the recorded SHD of algorithms A1, A2, A3, and A4 still have not been very good – namely, 14.6, 10.2, 10, and 9.8, respectively. With the data size of 5000 and 7500 (and the MCAR rate of 10%) the recorded average SHD of A1, A2, A3, and A4 are already much better – namely, 7.2, 4.2, 4.3, and 5.1, respectively. See Figure 9 for the learned models. With the data size of 10,000 we already get the true models except for the additional edges in the case of A2, as discussed in Section 4.2.3.

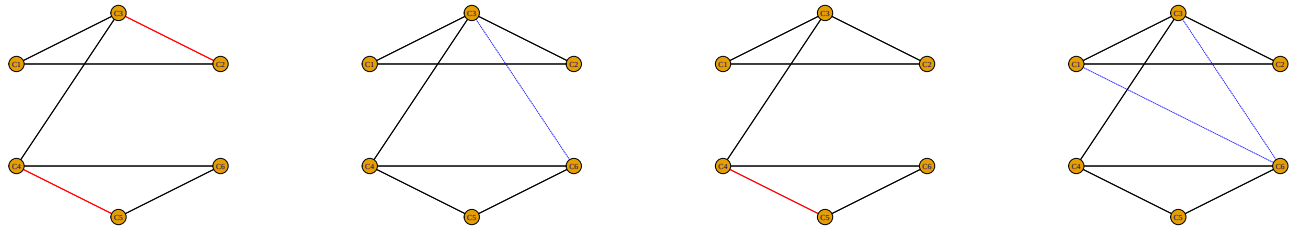


Figure 5: Models learned by A1, A2, A3, and A4, respectively, for most of ten datasets generated from the true N2 model (presented in Figure 11) (for $\alpha = 0$) with the MCAR rate 10 and the data size of 5,000.

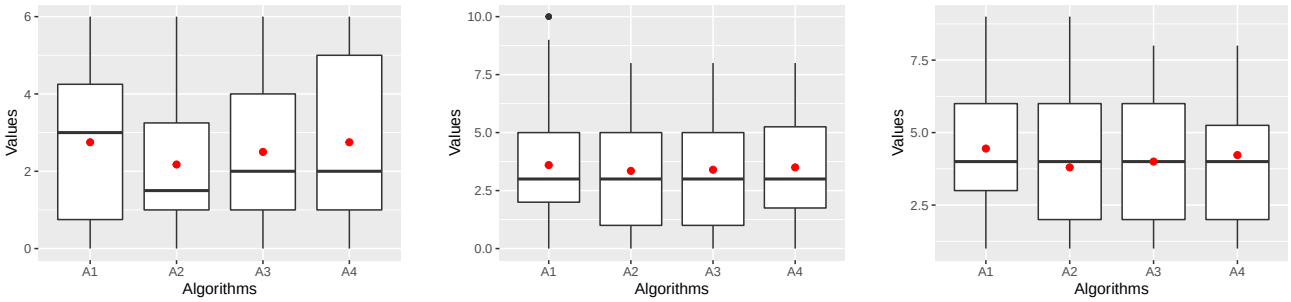


Figure 6: The Structural Hamming Distance to the true models of the resulting models of the structure learning algorithms using data generated from the BN2O model (the true model is presented in Figure 12) (with $\alpha = 0$) averaged over all data sizes for MCAR rates of 5%, 10%, and 15%, respectively.

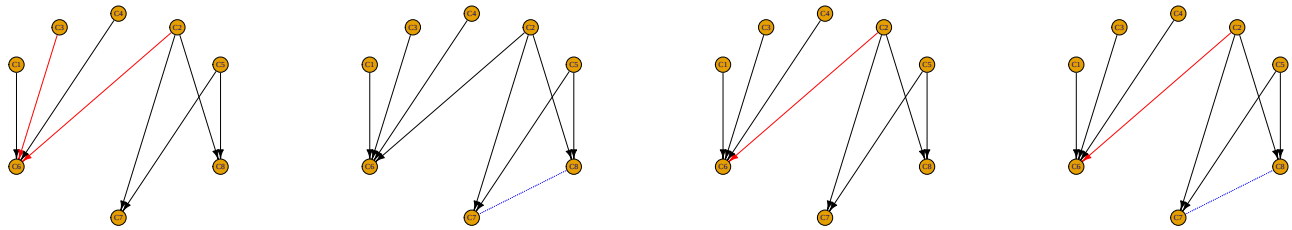


Figure 7: Models learned by A1, A2, A3, and A4, respectively, using data generated for most of ten datasets generated from the true BN2O model (presented in Figure 12) (for $\alpha = 0$) with the MCAR rate 10 and the data size of 5,000.

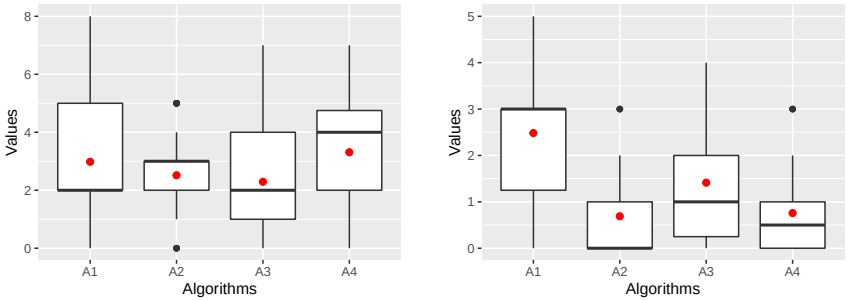


Figure 8: Results of the structure learning algorithms using data generated from the BN2O model (with $\alpha = 0$) with the data size of 3,000 and averaged over all tested MCAR rates. The plot on LHS displays the average number of additional edges and the plot on RHS displays the average number of missing edges.

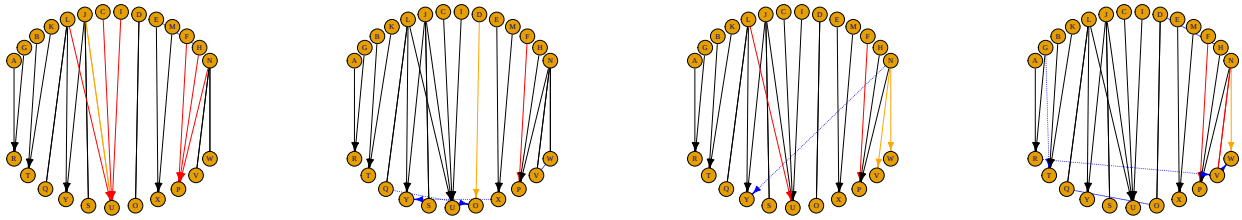


Figure 9: Models learned by A1, A2, A3, and A4, respectively, using the data generated from the large BN2O model consisting of 25 variables (for $\alpha = 0$) with the MCAR rate of 10% and the data size of 7,500 (true model is presented in Figure 13).

5 Conclusion

In this paper, we provide an approach to learning the optimal BN structure from incomplete data by adapting the considerations of [8]. This adaptation imputes missing values using product mixtures learned by the EM algorithm [11]. We have shown that the sequence of log-likelihood values generated by E-Step and M-Step of the EM algorithm is non-decreasing and the algorithm converges. Theorem 2.1 helps us reduce the collection of candidate parent sets for a variable, which can speed-up the learning algorithm.

We have performed experiments on incomplete data generated from different types of BN models to compare the proposed Algorithm A2 with other algorithms, namely with A1 [8], soft and hard EM [3], referred to as A3 and A4, respectively. In our comparisons, we use Structure Hamming Distance of CPDAGs of learned DAGs to CPDAGs of the original models.

Such comparisons have been undertaken on (a) general Bayesian networks and (b) Belief Noisy-or [4] (BNO) models with partially deterministic and nondeterministic conditional probability distributions. The experiments with models of type (b) are motivated by uncertain relationships in Bayesian networks, which are common in practical applications of BNs. We have obtained the following results in detailed simulation studies.

(a) General BN models:

- The A2 algorithm appears to be the best choice from among the tested algorithms for learning the structure of BNs from any incomplete data whatever the data size and the missing MCAR rate are.
- In most scenarios corresponding to different datasizes and MCAR rates, Algorithm A2 is significantly better than other algorithms and in no scenario is it significantly worse than any other algorithm according to the Wilcoxon test.

(b) BNO models:

- A2 is able to recover all true edges in the tested models except for the N1 model (shown in Figure 11) at size 1,000 and a missing rate of 15%. The different learned structure of the model N1 is acceptable because the Chi-square($\mathscr{X}^2$) test and the Kullback-Leibler distance (KLD) between the related conditional probabilities suggest there is a high degree of relationship between the connected variables.

- A2 has learned an additional edge in the case of Models N2 (shown in Figure 11) and BN2O (shown in Figure 12). The additional edge is acceptable since the $\mathscr{X}^2$ test and KLD suggest there is a high degree of relationship between these variables. We have seen that BIC of the learned structure is almost equal to BIC of the true model. For example, BIC of the model learned using A2 (shown in Figure 7) is -7,331.8 and BIC of the true model is -7,338.5. Similar behavior has been observed in other BNO models.

- A2 is always able to recover all edges while other algorithms are not.

- For large BN2O models, all algorithms require data sizes large than 3000 to have a good performance; e.g., for the BN2O with 25 variables A2 needs at least 10,000 data records to learn the correct model (with the exception of additional edges as discussed in Section 4.2.3).

We have empirically shown that our Algorithm A2 behaves better than other tested algorithms on several studied BNs and in different scenarios. Based on these experiments, we can recommend this algorithm for practitioners that use BNs or BNOs with incomplete data especially in the medical domain where BNO could be used to study the hidden relationship between symptoms and diseases. An interesting topic for future research might be learning the structure of large BN2O networks from incomplete data and optimize the number of components in the EM-Mixture .

Acknowledgement

This work was supported by Student Grant CTU SGS20/132/OHK4/2T/14

References

[1] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 20(2-3):131—163, 1997. URL: https://doi.org/10.1023/a:1007465528199.

[2] Cassio P de Campos, Mauro Scanagatta, Giorgio Corani, and Marco Zaffalon. Entropy-based pruning for learning Bayesian networks using BIC. *Artificial Intelligence*, 260:42—50, 2018. URL: https://doi.org/10.1016/j.artint.2018.04.002.

[3] Andrea Ruggieri, Francesco Stranieri, Fabio Stella, and Marco Scutari. Hard and soft EM in Bayesian network learning from incomplete data. *Algorithms*, 13(12):329, 2020. https://doi.org/10.3390/a13120329 doi:10.3390/a13120329.

[4] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[5] Nir Friedman and Moises Goldszmidt. Learning Bayesian networks with local structure. In *Learning in graphical models*, page 421—459. Springer, 1998. URL: https://doi.org/10.1007/978-94-011-5014-9_15.

[6] Zhifa Liu, Brandon Malone, and Changhe Yuan. Empirical evaluation of scoring functions for Bayesian network model selection. In *Proceedings of the Ninth Annual MCBIOS Conference. Dealing with the Omics Data Deluge*, Oxford, MS, USA., 2012. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-13-S15-S14 doi:10.1186/1471-2105-13-S15-S14.

[7] Poh Choo Song, Hui Yee Chong, Hong Choon Ong, and Sing Yan Looi. A model of Bayesian network analysis of the factors affecting student's higher level study decision: The private institution case. *journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(2):105—109, 2016.

[8] Cassio P de Campos, Zhi Zeng, and Qiang Ji. Structure learning of Bayesian networks using constraints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 113—120, New York, NY, USA, 2009. Association for Computing Machinery. https://doi.org/10.1145/1553374.1553389 doi:10.1145/1553374.1553389.

[9] James Cussens. Bayesian network learning with cutting planes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, page 153—160, Arlington, Virginia, USA, 2011. AUAI Press.

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1—38, 1977. URL: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[11] Jirí Grim, Jan Hora, Pavel Boček, Petr Somol, and Pavel Pudil. Statistical model of the 2001 Czech census for interactive presentation. *Journal of Official Statistics*, 26(4):673—694, 2010.

[12] J. Grim and P. Boček. Statistical model of prague households for interactive presentation of census data. In *SoftStat 95. Advances in Statistical Software 5. Conference on the Scientific Use of Statistical Software*, Heidelberg, DE, 1996.

[13] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289—317, 2016.

[14] Fred Glover. Tabu search-part I. *ORSA Journal on computing*, 1(3):190—206, 1989.

[15] Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks: with Examples in R*. Chapman & Hall, Boca Raton, 2014. URL: https://doi.org/10.1111/biom.12856.

[16] M Neuhäuser and Mann-Whitney Test. *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, 2011.

[17] Michael A Shwe, Blackford Middleton, David E Heckerman, Max Henrion, Eric J Horvitz, Harold P Lehmann, and Gregory F Cooper. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. *Methods of information in Medicine*, 30(04):241—255, 1991. URL: https://doi.org/10.1055/s-0038-1634846.

[18] B Abramson, J Brown, Ward E, Allan Murphy, and Robert L Winkler. Hailfinder: A bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1):57–71, 1996. Probability Judgmental Forecasting. URL: https://doi.org/10.1016/0169-2070(95)00664-8.

[19] A Philip Dawid. Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian statistics*, 4:109—125, 1992.

A Appendix A. Simulation Scenarios

This Appendix provides an inclusive list of all experiments in the simulation study described in Sections 3.2 and 4.2, organized by their main characteristics in Tables 8 and 9, respectively. The number of components in each experiment selected based on the number of variables in the datasets. The true models mentioned in the Table 8 are shown in Figure 10. The true models mentioned in the Table 9 are shown in Figures 11 and 12.

Table 8: Description of the key factors of all BN experiments in the simulation study.

Network	Missing Rate (MCAR)	Replicates	Sample Size
Weather [18]	10	10	100, 500,1000,5000,10000
	25	10	100, 500,1000,5000,10000
	50	10	100, 500,1000,5000,10000,13000
Child [19]	10	10	1000, 2000,3000,5000
	15	10	1000, 2000,3000,5000
	50	10	1000, 2000,3000,5000
M2 (Figure 10)	5	10	500,1000,1500,2500,5000
	10	10	500,1000,1500,2500,5000
	15	10	500,1000,1500,2500,5000
	25	10	500,1000,1500,2500,5000
M1 (Figure 10)	10	10	500,1500,2500,5000,10000,13000
	20	10	500,1500,2500,5000,10000,13000
	35	10	500,1500,2500,5000,10000,13000
	50	10	500,1500,2500,5000,10000,13000

Table 9: Description of the key factors of all Belief Noisy-OR experiments in the simulation study (true models are presented in Figures 11 and 12).

Network	Missing Rate (MCAR)	Replicates	Sample Size
BN2O	5	10	1000,1500,2500,5000
	10	10	1000,1500,2500,5000
	15	10	1000,1500,2500,5000
N1	5	10	1000,1500,2500,5000
	10	10	1000,1500,2500,5000
	15	10	1000,1500,2500,5000
N2	5	10	1000,1500,2500,5000
	10	10	1000,1500,2500,5000
	15	10	1000,1500,2500,5000
large BN2O	10	10	5000, 7500

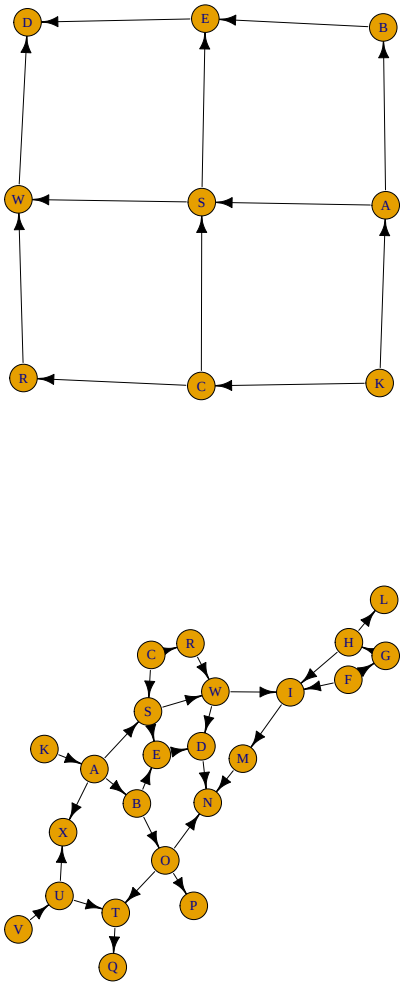
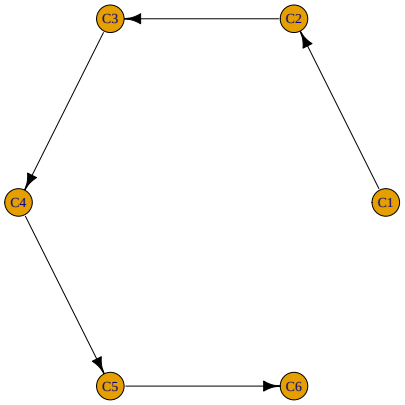


Figure 10: M1 and M2 true models, respectively



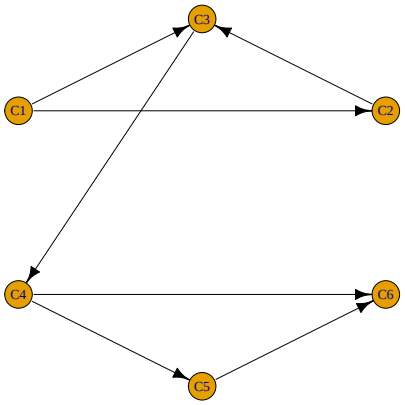


Figure 11: N1 and N2 true models, respectively. Their marginal probability distributions are summarized in Tables 5 and 6

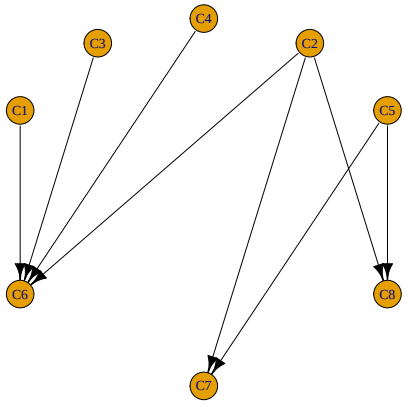


Figure 12: BN2O true model. Its marginal probability distributions are summarized in Table 7

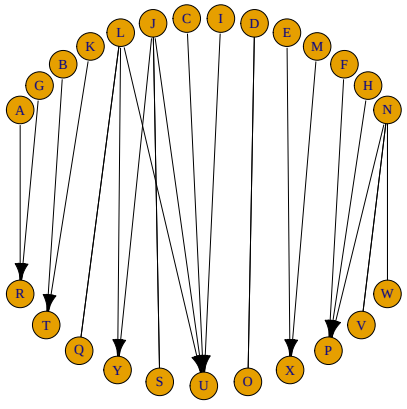


Figure 13: Example of a large BN2O model with 25 variables (whose learned models are presented in Figure 9).

A Prediction Model for Student Academic Performance Using Machine Learning

Harjinder Kaur[1],Tarandeep Kaur[1], Rachit Garg[2]
[1]Scool of Computer Applications, Lovely Professional University, Phagwara, 144401, India
[2]COS, School of Computer Science and Engineering, Lovely Professional University, Phagwara, 144401, India.
E-mail: Harjinder.12962@lpu.co.in, Tarandeep.24836@lpu.co.in, rachit.garg@lpu.co.in

Keywords: academic performance, decision tree education data mining, ensemble model, naïve bayes, performance prediction

Received: July 15, 2022

***Abstract**: Academic data mining impacts a large number of educational institutions, significantly, playing a prime role in accumulating, studying, and analyzing the academic data. The accumulated academic data can be processed and analyzed for various purposes. It can be used for predicting the student academic performance and thereby broadening the retention rate of academic institutions. The prediction of students' academic performance at the initial stage helps the students to identify their lacking subjects such that they can focus more on their deficient subjects and improvise their academic performance. Currently, numerous machine learning techniques are being used by the academic institutions to extract, analyze and predict the students' academic performance and identify the fast and slow learners. This paper proposes an ensemble model, using the voting method for preclusive prediction of the student academic performance. The predicted results are being further utilized by the poor performers to concentrate more on their deficit courses. Accordingly, the instructors can focus on creating and implementing novel strategies or amending the existing pedagogical tools and approaches to aid the slow learners in improvising their performance. The proposed model has been tested on the academic data of an educational institution using the RapidMiner tool. The results depicts that how the number of E grades proportionally affects the performance of the students in academics. The proposed ensemble model generates the predicted results with an accuracy of 90.83%.*

Povzetek: Predstavljena je metoda strojnega učenja za napovedovanje učnega uspeha.

1 Introduction

Academic Data Mining (ADM) has obtained astounding inquisitiveness in the recent years. The need for the analysis and assessment of the factors impacting the academic performance of students has embellished the demand for Academic Data Mining (ADM) or Educational Data Mining (EDM) [1]. Significantly, such factors can include student academic performance measured in terms of final grades obtained, course attendance, mid-assessment marks, etc. [2]. ADM plays a pivotal role in analyzing student performance based on the above-said factors and thereby classifying them into fast and slow learners. Additionally, ADM can also aid in providing subtle suggestions and recommendations for both the instructors as well as the students in improvising their performance. This can involve processes such as academic performance prediction and academic performance recommendations. Both the processes are essential for every educational institution as their reputation is centered upon the academic accomplishments of students [3]. The primary goal of academic performance prediction of learners is the identification of students at risk in their initial stage of career. This identification helps the instructor to analyze the factors affecting the performance such that corrective actions can be taken for the students at risk of lower achievement levels. Moreover, the timely analysis of weak performers benefits the academic institutions in increasing their retention rate [4].

The academic performance of students is predicted using different supervised learning techniques such as classification and prediction. Learning Analytics (LA) plays a very significant in the field of education. The motivation for using LA by academic institutions is to analyze the patterns obtained from the educational data after prediction. So, after the academic performance prediction, LA in association with ADM is used to generate effective results that leads to the categorization of different types of students [5].

This research proposes a model that serves as an alarming structure for educational organizations. The proposed model can be used by the students to discover and concentrate on their disconcerting subjects while the faculties can focus on improving their learning strategies towards such students. Currently, many machine learning algorithms are available for envisaging student educational performance and ADM [6, 7]. The proposed model is also an ensemble machine learning-based model that predicts the student's academic performance using an ensemble of machine learning algorithms,

Decision Tree, Naïve Bayes, and K-Nearest Neighbor. For performance prediction, the records are collected from the academic institution which is then pre-processed to eliminate anomalies so that only the data which is helpful for the analysis purpose is anomalies free. The cleaned data is then applied to the model and thereafter produced the predicted results.

1.1 Motivation for the work

Currently, the majority of the academic institutions face challenges related to the decreasing student academic performance and thereby rising student dropout ratio. This poses an alarming and stake-compromising situation for the academic institutions. They consistently struggle to maintain the retention rate of the students. Similarly, the decline in the student academic performance impacts a student physiologically, economically and socially. Some students get demotivated and resultantly think of discontinuing their degree. This leads to the increase in the dropout rate for the academic institution. Such circumstances are challenging for the teaching fraternity as well since the failure or decrease in the student academic performance puts a question mark on overall conduct of the teacher. It raises concerns on the teaching capabilities and pedagogical approach followed by him/her.

The proposed ensemble prediction model has been developed considering such circumstances. It helps in the reduction of drop-out rates and results in improving the retention rate of students. It provides the solution for the increase drop out issue faced by institutions by predicting the academic performance of the students precisely and proficiently. The proposed model has been trained using the historic data of students and then tested using the testing dataset. The predicted results classify the students into slow learners and fast learners. The proposed model serves as an alarming system for slow learners, the students who are at academic risk at the early stage of their carrier along with the courses affecting their performance. The early identification of students at academic risk helps the instructors to create new pedagogies, strategies and special academic counselling sessions for the weak students. Additionally, such initiatives helps the slow learners to concentrate more on their weak areas so that they can perform well in their academics and thereby improvising their performance. The improvement in the academic performance at early stage helps the slow learners to complete their degree on time that further improves the retention rate which further improves the repute of academic institutions.

Overall, the proposed model is useful for academic stakeholders including learners, instructors/ teachers and educational institutions. It benefits the learners in their self-assessment on academic background by providing the reasons which are responsible for their academic downfall. The model assist the instructors to keep track of the academic growth of the students and helps them to provide special attention towards the slow learners. The predicted results of the proposed model helps the educational institutions to devise new strategies and steps for promoting and educating slow learners for their performance improvement thereby increasing the retention rate of the institutions.

The rest of the paper has been divided into 5 sections. Section 2 lists a tabular representation of the existing techniques used for predicting students' academic performance. The proposed model has been elaborately discussed along with its structure and working in Section 3. Section 4 covers the empirical analysis of the proposed model on the collected data. The last fragment in the paper concludes with a brief description of why the ensemble approach has been preferred for predicting students' academic performance. It also concludes with an insight into the futuristic extensions that can be made in the proposed model.

2 Literature review

The existing educational research shows that the intersection of academic data and machine learning techniques is advantageous for carrying out interdisciplinary work [8]. Research on educational data helps in the identification and selection of various factors revealing argumentative and empirical academic results. The implementation of various machine learning techniques on collected academic records can help in developing dynamic alarming systems. Such systems will be beneficial for both instructor/tutor as well as learners to work in their lacking areas [9, 10].

Subsequently, the learners can improvise their academic performance based on the feedback of predicted results of alarming systems such that they can complete their respective degrees on time and with minimum dropouts or backlogs. Table. 1 illustrates the review of literature along with the techniques used and objectives of each model, and Figure. 1 shows the categorization of different prediction models based on machine learning.

Table 1: Existing academic performance prediction models.

Prediction Models	Machine Learning Technique(s) Used	Core Objective
[1]	Decision Trees, Support Vector Machines, Naive Bayes, Bagged Trees, and Boosted Trees	The early segmentation of students based upon their performance in the first year which helps in achievements of better results during the course completion.
[3]	Decision Trees	To categorize the students based upon their performance.
[4]	Logistic Regression, Neural Networks, Random Forests.	To identify the various challenges faced by the student in their first educational year based upon student registration data.
[5]	Decision Trees, Rule and Fuzzy Rule Induction Methods, and Neural Networks.	To predict the marks of university students in their final exams.
[11]	Logistic/Linear Regression, Matrix Factorization	To use educational data for an intelligent tutoring system.
[12]	Linear Regression, Neural Networks, Support Vector Machines	To predict the student score based upon their mid-term marks.
[13]	Neural Networks, Random Forests, and Decision Tree	To predict the student academic performance of first-year students
[14]	Linear regression, neural networks, support vector machines, decision trees, naive Bayes, k-nearest neighbor	To provide various courses based upon the existing data which help in improving the academic performance of a student.
[15]	Decision tree, Gradient boost algorithm, and Naïve Bayes	To identify the weak students and provide special counselling for their betterment.
[16]	SVM and Naïve Bayes	To predict the student's academic performance using Naïve Bayes and compare the predicted results with the results generated by SVM.
[17]	K-Nearest Neighbor, Naïve Bayes, Decision Tree, and Logistic Regression	The main objective of the study is to predict the student's academic performance along with the factors affecting their performance.
[18,19]	Decision Tree	To assess the student's academic performance using the decision tree. The predicted results were used to provide a recommendation to weak students so that they can improve their performance which lowers the failure rate.
[20]	Naïve Bayes, Neural Network, and Decision Tree	The main objective is this research is the usage of various data mining techniques to predict and analyse the academic performance of students founded from the academic data available by a participated forum.
[21,22]	Random Forest, Neural Networks, SVMs, and Regression Techniques	EDM was used to identify the weak students, based upon their performance. It also helps in the identification of various factors responsible for affecting and deteriorating the academic performance of the students.

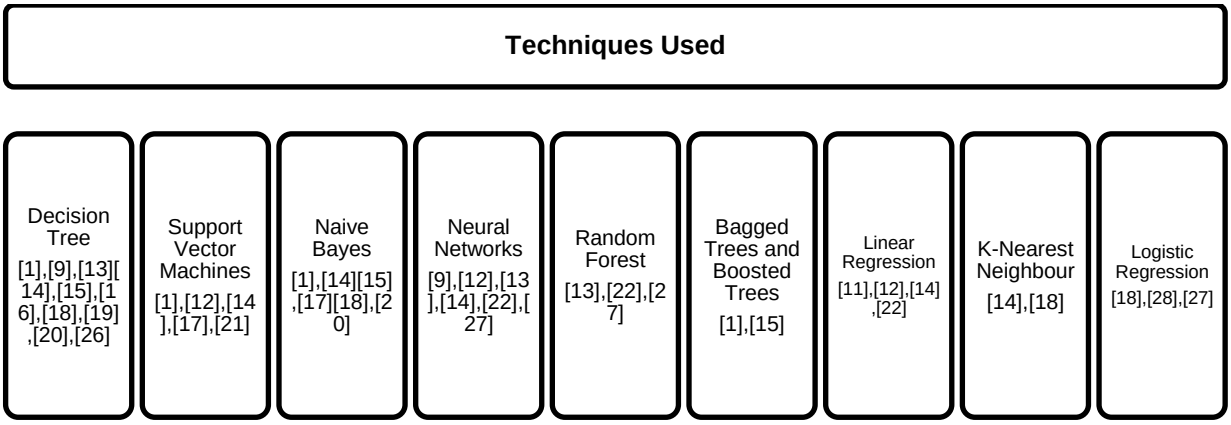


Figure 1: Categorization of existing prediction models based on the machine learning techniques used.

3 Proposed ensemble model

The primary goal of creating an ensemble model helps in the production of more accurate results as compared to the accuracy of results produced by individual classifier. The proposed model uses the ensemble of heterogeneous classifiers. The ensemble model proposed here accepts the output from multiple classifiers such as decision tree, Naïve Bayes, and K-NN. The proposed ensemble combines the output of heterogeneous classifies using voting approach which resultantly produces the final prediction results. The idea of ensemble approach works if and only if all the selected classifiers producing different class labels rather than agreeing on the same decision. Figure 2 depicts the flow of ensemble method.

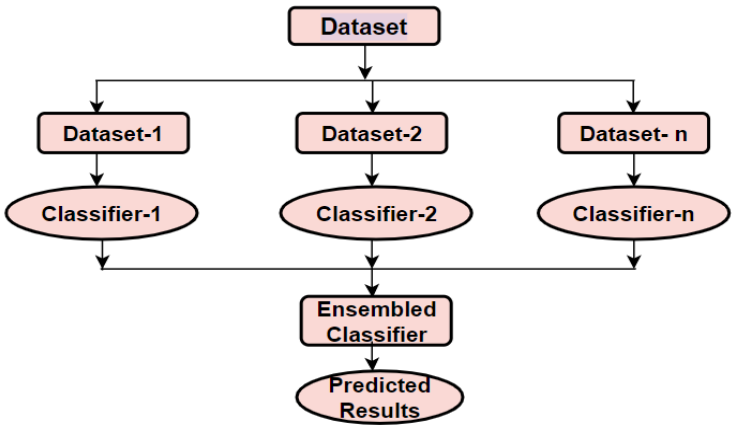


Figure 2: Basic ensemble approach for prediction.

The proposed ensemble model performs classification of the students based on their academic performance considering their marks in the courses inclusive of their attendance in each course. The data for classification has been collected from sources such as using Google form and a designed interface. Certain attributes generate irrelevant values such as incomplete data, duplicate data, naming identification problems and hence have no participation in the classification process. Thus, such irrelevant attributes were stricken out of the classification process else the use of these attributes could have increased the classification errors and complexity of the selected algorithm. Conclusively, this helped in making the predictions more accurate.

The proposed ensemble model has been designed to predict student academic performance using an ensemble of machine learning algorithms. The primary objective of designing an ensemble model is that every selected classifier must be complementary to each other in the context of a judgment so that further accuracy can be achieved [23]. The model intends to compute the student academic performance (in terms of Cumulative Grade Points) and achieve an early separation of learners

segregating them into slow and fast learners based upon their educational performance.

3.1 Working of the proposed ensemble model

When it comes to predicting student academic performance, a single classification model might not produce the appropriate outcome. Moreover, the single classification models suffer from high variance [24, 25].In the proposed ensemble approach, the output of multiple models has been combined which further enhances the overall accuracy of prediction results. There are some ensemble approaches like bagging, boosting, stacking, and voting with each having its pros and cons.

In the proposed model, voting technique has been used because the prediction results have been produced by combining the output of multiple classifiers. The results generated by the voting approach are better in comparison with a single classifier because in voting the decision depends upon the majority vote [26]. The choice of voting approach has been made because it produces predicted results with low variance in comparison to the variance produced by single classification model [27, 28].

The students are the key component of the proposed ensemble model as they provide their academic details as input. The academic details comprise their courses, marks/grade in each course, and attendance in individual courses as these academic parameters are considered as the crucial factors for measuring the academic performance of students. An interface has been designed to get the academic details of the students that are used for the model testing. The interface supports heterogeneous devices where the learners can provide their academic inputs by using either their smartphones, laptops, or even their desktops too.

The students input their educational details through the designed student interface. Such student academic data is stored in an academic database and is the core substantial asset for the prediction process. The stored data formulates different student records and is pre-processed, and then it is used to train the proposed model. During the pre-processing stage, the academic records have been integrated followed by checks to look for any inconsistencies, such as duplicates, missing values, etc. Consequently, the pre-processing stage generates the refined data which is further used to train the proposed ensemble model.

In the proposed ensemble model, the training dataset is used for the generation of rules which are being used for the prediction as shown in Figure 3. The testing dataset is being applied to constructed ensemble model to get the predicted academic performance based upon the rules generated using the training dataset.

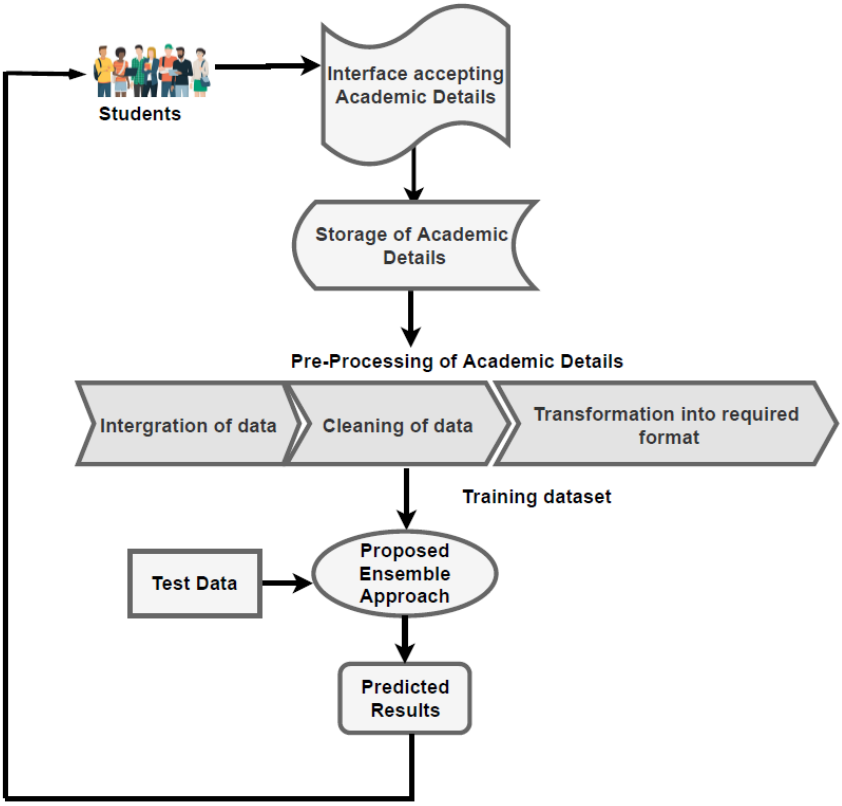


Figure 3: Proposed ensemble model.

The predicted results of the model are beneficial for both the instructor as well as the learner. It enables the instructor in scrutinizing the student's academic results and derive their performance from them which can be further used to take certain novel strategic actions for improvising the performance of slow learners. Concomitantly, this helps the recognizing the students at academic risk at the stage of academics which helps in augmenting the student retention rate and completion of degree on time. Also, the predicted academic performance is used as feedback by the students.

3.2 Mathematical formulation and proposed algorithm

Analytically, the proposed algorithm helps to categorize the different types of learners into strong and weak learners. The differentiation identifies the weak learners and also the courses in which they have underperformed. Subsequently, this helps the weak learners to concentrate more on such subjects they were lagging and resultantly improvise their performance. Identification of weak performers at early stage guides them to perform well in their end term exams. Mathematically, in order to categorize the students, their $CGPA$ has been calculated by considering their grade points and credit for each course. For calculating the $CGPA$, the student's grade points have been initially computed from the marks obtained in each course as shown in Table. 2.

The proposed model has based on certain assumptions which are as follows:

- The $CGPA$ of students has been calculated by considering the grade points of each course. In the proposed model for $CGPA$ calculation, the grade point consideration is at a 10 scale.
- The results of the proposed ensemble model used by 2nd-semester students further recommend the courses because in majority of the universities the selection option has been started from the second year onwards.
- The number of subjects considered for the calculation of $CGPA$ was 8.
- The total marks of various courses inclusive of attendance marks.
- For predicting the student academic performance the grade consideration is from A-E.

The following table shows the description of grade points and grades based upon the marks:

Table 2: Grade as per marks range.

Range of Marks	Grade Point	Grade
90 - 100	9.0 - 10.0	A+
80 - 89	8.0 - 8.9	A
70 - 79	7.0 - 7.9	B+
60 - 69	6.0 - 6.9	B
50 - 59	5.0 - 5.9	C
40 - 49	4.0 - 4.9	D
< 40	0.0-3.9	E

Objective Function: Map $\left(Stu_i, Cou_j, MC_{ij} \xrightarrow{yields} cgpa\right)$ (1)

Where:

Stu: Students

Cou: Courses

MC_{ij}: Marks obtained by i^{th} student in j^{th} course.

$CGPA$: Commulative Grade Point Assessment.

i: Index of Students$i \in S$ $where$ $S = \{1 \leq i \leq n\}$

$S = Set\ of\ students\ and\ n\ is\ the\ maxiumum\ number\ of\ students$

j: Index of Course and $j \in R$ $where$ $R = \{1 \leq j \leq m\}$

$R = Set\ of\ Courses\ and\ m\ is\ the\ maxiumum\ number\ of\ courses$

MCi, j: Marks in each course such that$i \in S$ and $j \in R$

$where\ S = \{1 \leq i \leq n\}\ and\ R = \{1 \leq j \leq m\}$

For accomplishing the objective function, a map function has been devised. The mapping function predicts the performance of the students by calculating their CGPA based upon the academic details given by students. Here, the map is the function that maps thei^{th} students in to their corresponding CGPA by considering

their course and marks in each course. The general formula for the calculation of $CGPA$ is depicted in Eq. (2).

$$\boxed{\text{CGPA} = \frac{\sum(\text{G}*\text{CR})}{\sum \text{CR}}} \quad (2)$$

Where:

$CGPA$ −Cumulative Grade point Average

CR −Represents the credit score of a course

G −Represents Grade points obtained by the student in a course.

The proposed model is composed of a set $St = \{stu_1, stu_2, st_{u3}, \ldots\ldots stu_n\}$ of n students such that $Stu = \{stu_i | 1 \leq i \leq n\}$ specifies the number of students; a set $Cou = \{cou_1, cou_2, cou_3, \ldots\ldots cou_m\}$ represents the m different subjects such that $Cou = \{cou_j | 1 \leq j \leq m\}$.

Let g_{ij}denotes the grade points obtained by the i^{th}student inj^{th}course. If $cgpa_i$ is the CGPA of i^{th} student, then it can be obtained by matrix algorithm specified in Eq. (3):

$$Cgpa_i = \begin{bmatrix} cgpa_1 \\ cgpa_2 \\ \vdots \\ cgpa_n \end{bmatrix} = \frac{1}{\sum_{j=1}^{m} cr_j} \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{bmatrix} \begin{bmatrix} cr_1 \\ cr_2 \\ \vdots \\ cr_m \end{bmatrix} \quad (3)$$

Where cr_j denotes the credits corresponding to j^{th}course$\forall\ 1 \leq j \leq m$, and the proposed algorithm is shown as follows:

Objective Function: Mapping of student with their CGPA by considering their program courses and marks in individual course which affect student's academic performance consists of {Students, Courses, Marks in each course}

Input: Student academic details

Output: Student categorization into weak and strong learners; Special inputs to weak students for improving their performance.

1. Perform preprocessing of collected data.
2. Use the pre-processed data as a training dataset.
3. Training dataset is used to train the model for the generation of rules.
4. Testing data is used for the prediction of performance using trained model;

   $\{stu_i, cou_j, MC_{ij}\}$ has been applied to map to get $cgpa_i$, using Eq. (1).

5. (a) Eq. (2) specifies the general formula for the calculation of $CGPA$.

   (b) $cgpa_i$ is computed using Eq. (3) where $cgpa_i$is the $CGPA$ of individual student.

6. The calculated$CGPA$ helps in the identification of weak and strong learners.
7. The predicted results are being used by the:

Learners (to improve their performance).

Instructors (to provide suggestive measures to poor performers)

4 Results

The experimental results have been obtained using the data from the department of computer science of an academic institution. The dataset contains 400 records of current students belonging to different sections of the computer science department. The dataset has been divided using the split operator, where 70% of the entire data is being used for training the model and the rest 30% is used for the testing of an ensemble model. Major attributes considered for analyzing the performance are the attendance in each course, the grade obtained in each course, the overall CGPA of the student, number of pending E grades. Figure 4 shows the results generated by the proposed ensemble model.

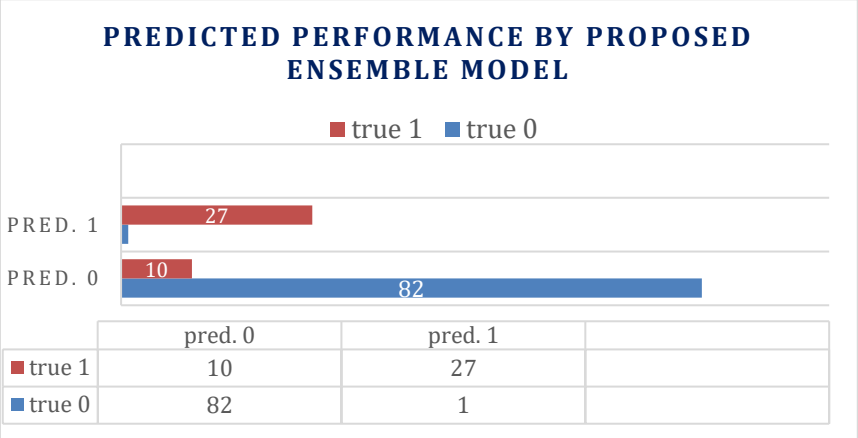


Figure 4: Results generated by ensemble method.

Performance vector shown in Tab. 3 proves the accuracy of the ensemble method using a vote operator that uses the majority vote from the base learners for predicting the results. The ensemble method has shown an accuracy of 90.83%. In the confusion matrix, 0 represents good performers and 1 denotes bad performers. For fast learners, 82 instances are correctly identified whereas 10 are incorrectly identified. Similarly, for bad performers, 27 instances are correctly identified whereas 1 is incorrectly identified.

Table 3: Performance vector (ensemble method).

	true 0	true 1	class precision
pred. 0	82	10	89.13%
pred. 1	1	27	96.43%
class recall	98.80%	72.97%	

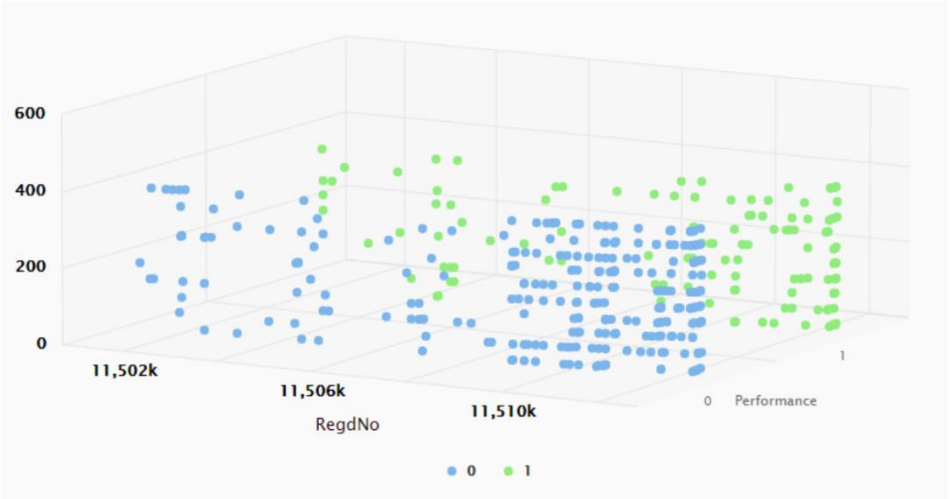


Figure 5: Relationship between actual and predicted performance.

The scattered 3D plot view of the relationship between actual and predicted results generated by the proposed ensemble model is illustrated in Figure 5 where x axis represents the RegdNo and the value column signifies performance in terms of slow (1) and fast learners (0).

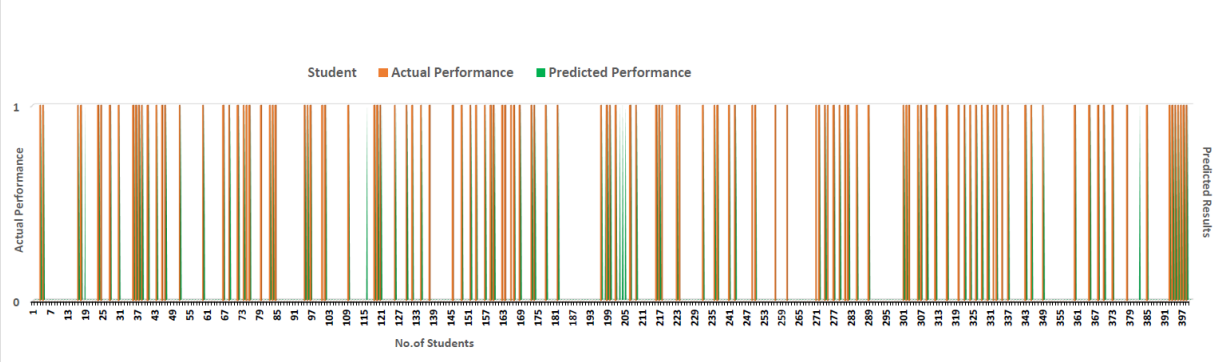


Figure. 6: Actual and predicted results based on E grades.

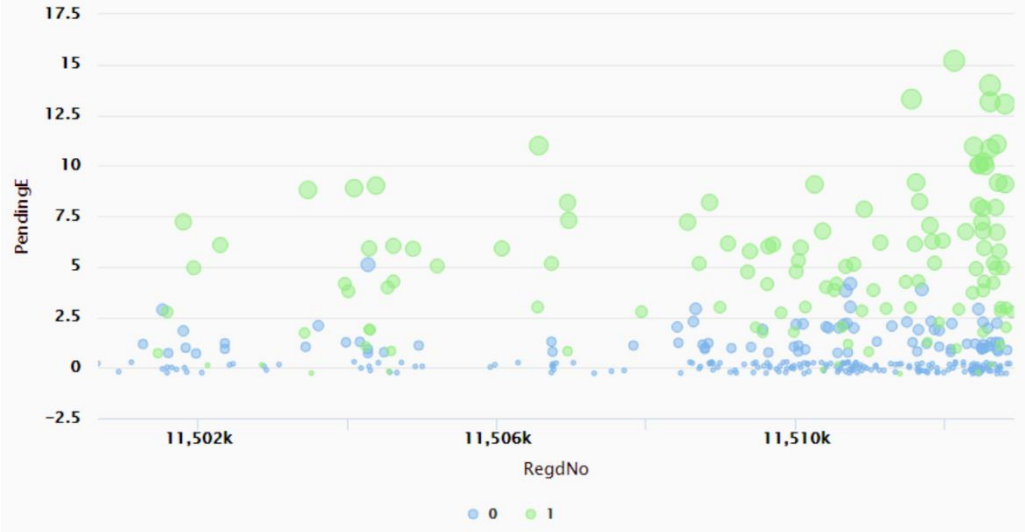


Figure 7: Predicted performance based upon pending E grades.

The actual and predicted results based on E-grades has been depicted in Figure 6 and 7. Figure 8 illustrates the registration-wise predicted performance of students after the re-appear exam has been given. The blue colour circle indicates good performance that is signified by 0 whereas the green colour represents the poor performance of the student which is denoted using 1. The results show that the more the number of re-appears a student is having considered under the category of a poor performer. Therefore, corrective actions for such students are required to be taken on time by the student as well as from the instructor.

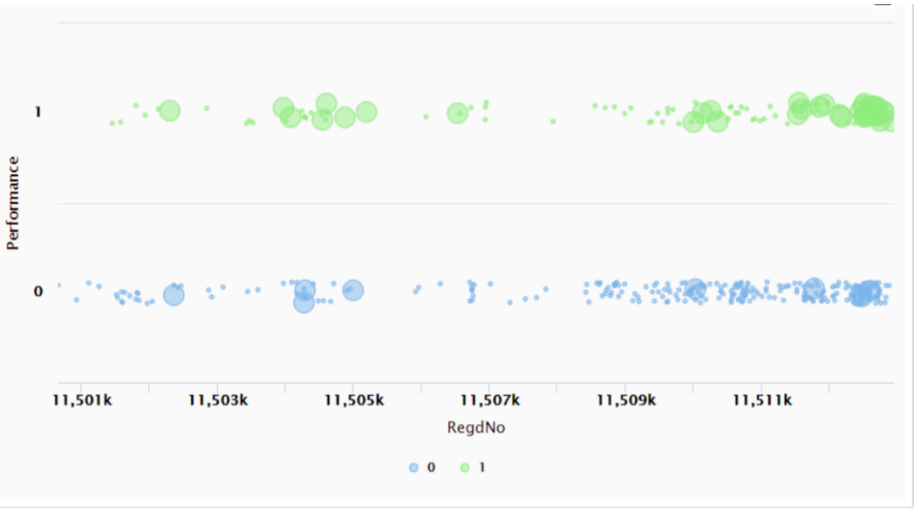


Figure 8: Predicted results considering reappear exam given.

5 Conclusion and future directions

Presently, the academic educational institutions are facing difficulty in sustaining the low retention rate of students. The task of maintain the retention rate can only be achieved by reducing the drop-out ratio of students. The high student retention rate depends significantly on the student academic performance. It becomes highly important for the academic institutions to predict the student performance for subsequent sessions such that retention rate can be maintained as well student performance can be improved. Also, the prediction of student academic performance at an early phase of their degree helps to do self-assessment for their downfall so that the student can do the corrective actions to improvise his/her on time. The model is helpful for the instructors as well who can verify and revise their pedagogical approaches if required.

A lot of research is being carried out to develop models for predicting the student academic performance using academic data mining strategies. Various machine learning techniques have been used to develop such predication models that act as an aid for the academic institutions. The paper proposes an ensemble model based on machine learning techniques, Decision Tree, Naïve Bayes, and K-NN classification algorithms catering to such problems. It helps in identifying the weak learners by predicting their performance based upon the historical academic data. The model has been implemented on a gathered dataset and achieves an accuracy 90.83%.

The research work presented in this paper can be further extended to develop a recommender system that will use the performance prediction results and subsequently recommend course-specific elective courses to the students. Such recommendations tend to augment student skills depending on their performance. Additionally, a recommender system can be developed that offers students interest-oriented or choice-driven suggestions regarding course selection considering and mapping the student's previous performance along with the student choice. The major research for the academic performance prediction of the students considers the direct factors (such as courses, marks in each course, attendance and grades etc.). The incorporation of the indirect factors (such as physiological, behavioral, economic and social etc.) that affect the student academic performance can be carried out further.

Recently, several edtech companies have emerged during COVID 2019 era. Such companies are engaged in the practice of incorporating Information Technology (IT) and digital tools for the student learning and engagement. The edtech companies are now using predictive analytics for mining student academic records, enrollment, attendance, class engagement, etc. The edtech companies can use the prediction as well as recommendation models to help the students by suggesting the appropriate course based upon their predicted performance.

Acknowledgement: Mohamed Alwanin would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-###. The authors deeply acknowledge the Researchers Supporting Program (TUMA-Project-2021-14), AlMaarefa University, Riyadh, Saudi Arabia for supporting steps of this work.

***Funding Statement*:** Mohamed Alwanin like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-###. This research was supported by Researchers Supporting Program (TUMA-Project-2021-14), AlMaarefa University, Riyadh, Saudi Arabia.

Conflicts of Interest: Authors declare that there is no conflict of interest associated with this study.

References

[1] V.L. Miguéi, A. Freitas, P.J.V. Garcia and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decision Support System*,vol. 6, no. 5, pp. 65-78, 2018.

[2] S. J. Lakshmi and M. Thangaraj, "Recommender system for stimulating the learning skill of slow learner in higher educational institution using EDM," International Journal on Recent Technolofical Engineering, vol. 5, pp. 98-109, 2019.

[3] D. T. Ha, P. T. T. Loan, C. N. Giap and N. T. L. Huong, "An empirical study for student academic performance prediction using machine learning techniques," International Journal of Computer Science and Information Security (IJCSIS), vol. 18, no. 3, pp. 75-82, 2020

[4] R. Umer, T. Susnjak, A. Mathrani and S. Suriadi,"On predicting academic performance with process mining in learning analytics," *Journal of Resource Innovation and Teach Learnearning*, vol. 78, pp. 155-168, 2017.

[5] O.H.T. Lu, A.Y.Q. Huang, J.C.H. Huang, A.J.Q Lin, H. Ogata *et al.*, "Applying learning analytics for the early prediction of students' academic performance in blended learning," *Educational Technological Socoety*,vol. 55, pp. 111-123, 2018.

[6] O. Viberg, M. Hatakka, O. Bälter, A. Mavroudi,"The current landscape of learning analytics in higher education,"*Computers in Human Behavior*, vol. 18, pp. 1001-1222, 2018.

[7] M. S. B. M. Azmi and I. H. B. M. Paris, "Academic performance prediction based on voting technique," in *2011 IEEE 3rd International Conference on Communication Software and Networks* , Calcuta, India, pp. 24-27, 2011.

[8] Tarandeep Kaur, Harjinder Kaur, "Machine Learning: An Internal Review", *Journal of Emerging Technologies and Innovative Research*, 5, no. 11, 6, 2018.

[9] C. Romero, P.G. Espejo, A. Zafra, J.R. Romero and S. Ventura, "Web usage mining for predicting final marks of students that use Moodle courses," *Computer Application in Engineering and Education*,vol. 65, pp. 555-578, 2013.

[10] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414-422, 2015.

[11] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 20, pp. 55-65, 2010.

[12] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models,"*Comput Education*, vol. 55, no. 6, pp. 33-42, 2013.

[13] M. Imran, S. Latif, D. Mehmood and M. S. Shah,"Student academic performance prediction using supervised learning techniques," *International Journal on Emerging Technologies in Learning*, vol. 77, pp. 102-120, 2019.

[14] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira and R. Abreu,"A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance,", in *Proc. ICEDM*, Noida, India, pp. 55-64, 2015.

[15] P. Kamal and S. Ahuja,"An ensemble-based model for prediction of academic performance of students in undergrad professional course," *Journal of Engineering Design and Technology*, vol. 98, pp. 654-672, 2019.

[16] V. Skrbinjek and V. Dermol, "Predicting students' satisfaction using a decision tree,"*Tert Education and Management*,vol. 64, pp. 210-218, 2019.

[17] Dr. Antino Marelino. (2014). Customer Satisfaction Analysis based on Customer Relationship Management. International Journal of New Practices in Management and Engineering, 3(01), 07 - 12. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/26

[18] Dr. Sandip Kadam. (2014). An Experimental Analysis on performance of Content Management Tools in an Organization. International Journal of New Practices in Management and Engineering, 3(02), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/27

[19] Ms. Nora Zilam Runera. (2014). Performance Analysis on Knowledge Management System on Project Management. International Journal of New Practices in Management and Engineering, 3(02), 08 - 13. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/28

[20] Mrs. Leena Rathi. (2014). Ancient Vedic Multiplication Based Optimized High Speed Arithmetic Logic. International Journal of New Practices in Management and Engineering, 3(03), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/29Kaur H, Kushwaha AS., "An elicit elucidation on the process of education data mining" , International Conference on Intelligent Computing and Control Systems, ICCS 2019.

[21] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, Korat, Thailand, pp. *568-572*, 2017.

[22] S. Poonam, S. Ahuja, V. Jaitly and S. Jain, "A framework to alleviate common problems from

recommender system,"A case study for technical course recommendation," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no.2, pp. 451-460, 2020.

[23] A. Rajak, A. K. Shrivastava and V. Vidushi, "Applying and comparing machine learning classification algorithms for predicting the results of students," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no.2, pp. 419-427, 2020.

[24] H. Guruler, A. Istanbullu and M. Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases," *Computer Education*, vol. 6, no. 5, pp. 125-138, 2010.

[25] A. Rajak, A. K. Shrivastava and V. Vidushi, "Applying and comparing machine learning classification algorithms for predicting the results of students," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no.2, pp. 419-427, 2020.

[26] A. Siddique, A. Jan, F. Majeed, A.I. Qahmash, N.N. Quadri *et al.*, "Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers," *Applied Sciences*, vol. 11, no. 24, pp. 11845, 2021.

[27] A. S. Hoffait and M. Schyns,"Early detection of university students with potential difficulties," *Decision Support System*, vol. 9, no. 5, pp. 5-20, 2017.

A Multi-channel Convolutional Neural Network for Multilabel Sentiment Classification Using Abilify Oral User Reviews

Tina Esther Trueman[1], Ashok Kumar Jayaraman*[,2], Jasmine S[2], Gayathri Ananthakrishnan[3] and Narayanasamy P[4]
[1]Department of Computer Science, University of the People, Pasadena, United States
[2]Department of Information Science and Technology, Anna University, Chennai, India
[3]Department of Information Technology, Vellore Institute of Technology, Vellore, India
[4]Department of Electrical and Electronics Engineering, PSG College of Technology, Coimbatore, India
E-mail: tina.trueman@uopeople.edu, jashokkumar83@auist.net, jasminemtech7@gmail.com, gayathri.a@vit.ac.in, drpnsam@gmail.com
*Corresponding author

Keywords: Multilabel classification, sentiment classification, multichannel convolutional neural network, abilify user reviews

Received: April 13, 2021

Nowadays, patients and caregivers have become very active in social media. They are sharing a lot of information about their medication and drugs in terms of posts or comments. Therefore, sentiment analysis plays an active role to compute those posts or comments. However, each post is associated with multilabel such as ease of use, effectiveness, and satisfaction. To solve this kind of problem, we propose a multichannel convolution neural network for multilabel sentiment classification using Abilify oral user comments. The multichannel represents the multiple versions of the standard model with different strides. Specifically, we use the pre-trained model to generate word vectors. The proposed model is evaluated with multilabel metrics. The results indicate that the proposed multichannel convolutional network model outperforms the traditional machine learning algorithms.

Povzetek: Razvita je konvolucijska mreža za preučevanje izmenjav mnenj o boleznih na socialnih omrežjih.

1 Introduction

Social media has become an active part of drugs and medication users. They share the advantage or disadvantages of their medication and drugs. This information may give some insightful information about the reaction of the drug. Therefore, sentiment analysis plays a wide role to compute the opinions of drug users and caregivers. The sentiment analysis can be performed at the document level, sentence level, or aspect level [1, 2]. The document and sentence level computes the overall opinion. But, the aspect level computes opinion at a specific target or an entity. In this paper, we aim to focus on aspect level sentiment. A comment may be associated with a single label or multilabel [3]. The single label problem has only one label. However, It has two classification methods namely, binary classification or multiclass classification [4]. The binary classification problem belongs to a binary set such as true and false or positive and negative. The multiclass classification problem belongs to a set of more than two elements such as positive, neutral, and negative. In these problems, algorithms assign only one label to comment or instance. Multilabel classification problem belongs to a set of multiple target labels where each label maybe belongs to a binary class or multiclass.

Traditionally, the multilabel classification problems are solved using problem transformation, adapted algorithms, and ensemble learnings [3]. The problem transformation problem is further solved using the binary relevance, classifier chain, and label powerset methods [5, 6]. However, these methods use the traditional bag of words (BoW) method to represent features. These features fail to represent semantic meaning between words. Therefore, deep learning models are proposed to capture the semantic meaning between words in the input sequence. It is also proven that they outperform in many tasks such as image classification, text classification, etc [7, 8, 9]. In this paper, we propose a multichannel convolution neural network for multilabel sentiment classification using Abilify oral user comments. The multichannel model represents the multiple versions of the standard model with different strides. Particularly, we use the GloVe pre-trained model [10] to generate word vectors. We then evaluate the proposed multilabel metrics.

This paper is organized as follows. Section 2 briefly describes the related works. The proposed multichannel convolutional neural network for multilabel sentiment classification is presented in Section 3. In Section 4, the results and their comparison is presented. Finally, Section 5 concludes the paper.

2 Related works

In recent years, researchers widely studied clinical text and user text using natural language processing (NLP). They used both machine learning and deep learning to solve their problems. In this paper, we present the existing works on biomedical texts. Baumel et al. [11] investigated four models such as SVM, CNN, CBOW, and hierarchical attention-based recurrent neural network models for the extreme multilabel task using the MIMIC datasets. The authors indicated that the hierarchical attention-based recurrent neural network model achieves a 55.86% F1 score. Wang et al. [12] developed a rule-based algorithm to generate labels that are weakly supervised. Then, the authors used the pre-trained word embeddings to represent deep features. They employed SVM, random forest, MLPNN, and CNN algorithms. Their study indicated that the CNN model achieves the best performance score. Singh et al. [13] developed an attentive neural tree decoding model for tagging structured bio-medical texts with multilabel. This method decodes an input sequence into a tree of labels. The authors suggested that the proposed model outperforms on SOTA (sate-of-the art) approaches with biomedical abstracts. Citrome [14] reviewed the treatment of Abilify oral users with bipolar I disorder and schizophrenia. The author indicated that the tolerability of Abilify with schizophrenia appears superior to haloperidol, risperidone, and perphenazine. Rios et al. [15] demonstrated the biomedical text classification task using CNN. They indicated that they achieved a 3% improvement over the SOTA results.

Moreover, Gargiulo et al. [16] presented a deep neural network (DNN) for extreme multilabel and multiclass text classification tasks. The authors used two models: the first one uses a word embedding with two dense layers, and the second uses the convolution, word embedding, and the dense layers. Kolesov et al. [17] performed multilabel classification on incompletely labeled biomedical texts using the SVM and RF. They used soft supervised learning and weighted k-nearest neighbor algorithms for modifying the training set. Their study indicated that both algorithms perform better. Parwez et al. [18] presented the CNN model for multilabel text classification. The authors used the domain-specific and generic based pre-trained model to predict class labels. In summary, the above authors used SVM, NB, RF, and CNN to perform multilabel classification tasks on various biomedical texts (Table 1). In this paper, we propose multichannel convolutional neural network for multilabel sentiment classification using Abilify oral user comments.

3 The proposed method

In this section, we present a multichannel convolutional neural network for a multilabel sentiment classification model using Abilify oral user comments. The system architecture is shown in Fig.1. It includes data pre-processing, word embedding, multichannel CNN, merge layer, fully connected layer, and an output layer. Each of these processes is explained as follows.

3.1 Abilify oral dataset

We obtained this Abilify oral dataset from the IEEE Dataport [19, 20]. It contains 1722 user comments with their age group, gender, treatment condition, patient type, treatment duration, and labeled sentiment on satisfaction, effectiveness, and ease of use.

3.2 Pre-processing

The dataset is converted from upper case to lower case, removed punctuation lists and stop words, and retained the numbers where it describes the drugs in grams. Then, each instance is split into separate words using the tokenization method.

3.3 Multichannel convolutional neural network

The multichannel convolutional neural network represents the multiple version of the standard convolutional neural network model with different sizes of kernels. This representation allows the instance or document to process in different n-grams such as 4-gram, 6-gram, and 8-grams at the same time [22]. In particular, we define the standard convolutional neural network model with a word embedding layer, one-dimensional convolutional layer, dropout layer, max-pooling, and flatten layer. This standard version is defined with three channels for different n-grams. Each component of the channel is explained as follows.

3.3.1 Word embedding

In NLP, word embedding represents a feature learning technique where it maps the vocabulary of words or phrases into a vector space. Specifically, we use the GloVe word embedding [10] technique to generate word vectors in a fixed dimension with the semantic relationship between words.

3.3.2 Convolutional layer

Convolutional neural networks perform well in image classification and computer vision-related tasks. The convolutional layer is an important part of the convolutional neural network. It slides over an input sequence with a fixed kernel size to generate feature maps [15, 16, 18, 22, 23]. In this work, we use one-dimensional convolutional layers to move the kernel in one direction. This layer is mostly used to perform NLP tasks. The input and output of the 1D convolutional layer are 2D. The convoluted feature maps output the maximum, minimum, or average values using pooling layers.

Authors	Dataset	Models	Accuracy	Key Findings
Baumel et al. [9]	MIMIC Datasets	HA-GRU	55.86%	Classification of patient notes on ICD code assignment
Wang et al. [10]	Mayo Clinic smoking status	CNN	92.00%	A rule-based algorithm to generate labels that are weakly supervised
Singh et al. [11]	Articles describing randomized controlled trials	NTD-s	32.70%	An attentive neural tree decoding model for tagging structured bio-medical texts with multilabel
Rios et al. [13]	MED-LINE Citations	CNN-Vote2	64.69%	Biomedical text classification
Gargiulo et al. [14]	PubMed Dataset	CNN-Dense	20.15%	Extreme multilabel and multiclass text classification tasks
Kolesov et al. [15]	AgingPortfolio Dataset	SVM	30.59%	Multilabel classification on incompletely labeled biomedical texts
Parwez et al. [16]	Tweets dataset	CNN-PubMed	94.12%	Domain-specific and generic based pre-trained model to predict class labels

Table 1: Summary of the related works.

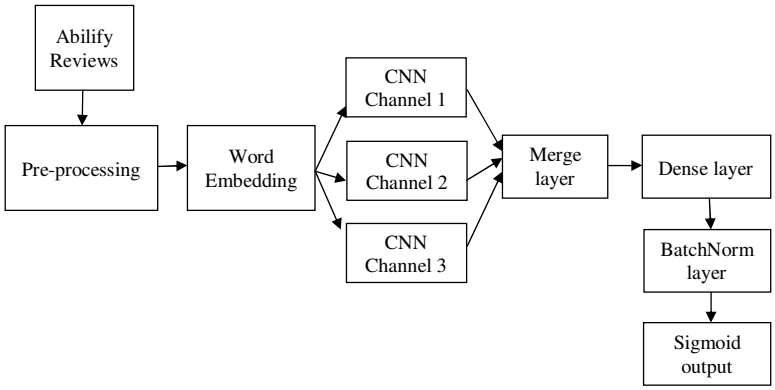


Figure 1: A multichannel convolutional neural network model.

3.3.3 Dropout layer

This layer is used to regularize the neural networks in terms of overfitting and underfitting. Specifically, it ignores some of the outputs in the neural network during the training process.

3.3.4 Max-pooling

The max-pooling layer is applied over each feature map to select the maximum value based on the filter size. It is smaller in size than the feature map. The output of this layer contains the most important feature values of the previous feature map [15, 16, 18, 22].

3.3.5 Flatten layer

The flatten layer converts the pooled feature map into a single column or one-dimensional array. This result is passed to a merged layer.

3.4 Merge layer

The merged or concatenate layer combines the output of each channel. These combined results passed to a fully connected or dense layer.

3.5 Fully connected layer

A fully connected or dense layer connects the input of the flatten layer to all units of the next layer. It works the same as the feed-forward neural network.

3.6 Batch normalization layer

The batch normalization layer allows all layers of a network to learn more independently. Specifically, it standardizes or normalizes the result of previous layers. Also, this layer acts as a regularization parameter to avoid overfitting.

3.7 Sigmoid output layer

The sigmoid output function predicts the probability-based output for each label as shown in equation 1. It is successfully applied in multilabel classification problems [24].

$$f(x) = \frac{1}{1 + exp^{x}} \tag{1}$$

4 Results and discussion

We implemented the proposed multilabel multichannel model on Abilify oral dataset. This dataset contains 1722 instances associated with a set of labels, namely, ease of use, satisfaction, and effectiveness. We split the dataset into training (1394), validation (155), and testing (173). The data instances are preprocessed with various tasks such as removing punctuations, stopwords, upper case to lower case, and tokenization. Then, word vectors are generated to the input sequences using the GloVe word embedding model. The proposed multichannel convolutional network model was applied to this dataset to perform the multilabel classification task. This model represents three multiple versions of the standard convolutional neural network with different kernel sizes. The standard CNN model consists of a word embedding layer, 1D-convolutional layer, dropout layer, max-pooling layer, and a flattened layer. The output

Data	Accuracy score	Hamming loss	F1 micro score	Accuracy per label		
				0	1	2
Validation	0.548	0.275	0.839	0.820	0.931	0.750
Testing	0.538	0.303	0.820	0.815	0.912	0.715

Table 2: Performance of the proposed multichannel CNN model.

Data	Accuracy score	Hamming loss	F1 micro score	Accuracy per label		
				0	1	2
BR_NB	55.2	0.379	71.5	60.5	68.9	57.0
BR_DT	59.6	0.343	73.9	61.9	79.7	55.3
BR_SVM	60.7	0.338	75.2	62.6	78.3	57.8
CC_NB	55.0	0.378	71.5	60.5	68.7	57.5
CC_DT	61.0	0.347	74.6	61.9	78.3	51.5
CC_SVM	60.5	0.346	75.0	62.6	77.5	55.7
LP_NB	54.1	0.385	69.4	566.	72.0	55.7
LP_DT	60.3	0.354	74.5	60.9	77.7	55.1
LP_SVM	**62.8**	0.334	76.5	63.4	79.8	56.4
Proposed	53.8	**0.303**	**82.0**	**81.5**	**91.2**	**71.5**

Table 3: Model comparison.

of each channel is combined through a merged layer and it is passed to a dense layer, batch normalization layer, and the sigmoid output layer. Specifically, we fixed the following hyperparameters using the random approach such as input length with 150 units, 100 embedding dimension, three kernel sizes (4, 6, and 8), ReLU activation, 0.8 dropouts, pooling size 2, 10 units in the fully connected layer, 20 epochs, and Adam optimizer with a binary cross-entropy loss function. The proposed multichannel CNN model for multilabel classification is evaluated using various multilabel metrics, namely, accuracy or exact match, hamming loss, F1-micro average score, and accuracy per label [3, 5, 20, 21]. Table 2 shows the performance of the proposed multichannel CNN model for multilabel classification. This result is compared with the problem transformation approaches, namely, binary relevance, label powerset, and classifier chains with NB, DT, and SVM [20] as shown in Table 3. The existing researchers in the Table 1 have addressed the multilabel classification using different biomedical texts. In this work, we used the patients and caregivers' opinion on drugs and medications dataset. In particular, we have compared the results of our proposed method with various baselines as shown in Table 3. The proposed multichannel CNN model achieves better results in terms of Hamming loss (30.3%), F1 micro score (82.0%), and accuracy per label (81.5%, 91.2%, 71.5%).

5 Conclusion

In this paper, we proposed a multichannel convolution neural network for multilabel sentiment classification using Abilify oral user comments. A pre-trained model was used to generate word vectors. Then, the proposed model was evaluated with the multilabel classification metrics. The results showed that the proposed multichannel CNN model achieves the better result in terms of Hamming loss, F1 micro score, and accuracy per label than the problem transformation approaches. In future work, we study the trend of drugs and medications in different age groups using patient and caregiver reviews.

Acknowledgement

We thank the Department of Information Science and Technology, Anna University, Chennai for the facility provided during this work.

References

[1] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82-89, 2013. https://doi.org/10.1145/2436256.2436274

[2] Bing Liu. Sentiment analysis: Mining opinions, sentiments, and emotions. *Cambridge university press*, 2020. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis-tutorial-2012.pdf

[3] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1-13, 2007. https://doi.org/10.4018/978-1-59904-951-9.ch006

[4] Sadri Alija, Edmond Beqiri, Alaa Sahl Gaafar, Alaa Khalaf Hamoud. Predicting Students Performance

Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection. *Informatica*, 47(1), 2022. https://doi.org/10.31449/inf.v47i1.4519

[5] Read J, Pfahringer B, Holmes G, and Frank E. Classifier chains for multi-label classification. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 254-269, 2009. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04174-7_17

[6] Tsoumakas G, Katakis I, and Vlahavas I. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079-1089, 2010. https://doi.org/10.1109/TKDE.2010.164

[7] LeCun Y, Bengio Y, and Hinton G. Deep learning. *Nature*, 521(7553):436-444, 2015. https://doi.org/10.1038/nature14539

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. *Cambridge: MIT press*, 1(2), 2016. http://www.deeplearningbook.org

[9] Patel R, Tanwani S, and Patidar C. Relation Extraction Between Medical Entities Using Deep Learning Approach. *Informatica*, 45(3), 2021. https://doi.org/10.31449/inf.v45i3.3056

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543, 2014. https://doi.org/10.3115/v1/d14-1162

[11] Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, and Elhadad N. Multi-label classification of patient notes a case study on ICD code assignment. 2017. https://arxiv.org/abs/1709.09587

[12] Wang Y, Sohn S, and Liu S et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, 19(1):1, 2019. https://doi.org/10.1186/s12911-018-0723-6

[13] Singh G, Thomas J, Marshall IJ, Shawe-Taylor J, and Wallace BC. Structured multi-label biomedical text tagging via attentive neural tree decoding. 2018. https://arxiv.org/abs/1810.01468

[14] Leslie Citrome. A review of aripiprazole in the treatment of patients with schizophrenia or bipolar I disorder. *Neuropsychiatric Disease and Treatment*, 2(4):427, 2006. https://doi.org/10.2147/nedt.2006.2.4.427

[15] Rios A and Kavuluru R. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. *In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, 258-267, 2015. https://doi.org/10.1145/2808719.2808746

[16] Gargiulo F, Silvestri S, and Ciampi M. Deep Convolution Neural Network for Extreme Multi-label Text Classification. *In HEALTHINF*, 641-650, 2018.

[17] Kolesov A, Kamyshenkov D, Litovchenko M, Smekalova E, Golovizin A, and Zhavoronkov A. On multilabel classification methods of incompletely labeled biomedical text data. *Computational and mathematical methods in medicine*, 2014. https://doi.org/10.1155/2014/781807

[18] Parwez MA and Abulaish M. Multi-label classification of microblogging texts using convolution neural network. *IEEE Access*, 7:68678-68691, 2019. https://doi.org/10.1109/ACCESS.2019.2919494

[19] Ashok Kumar J, Abirami S, and Tina Esther Trueman. Abilify Oral user reviews. *IEEE Dataport*, 2020. https://dx.doi.org/10.21227/p1jp-2m84

[20] Kumar JA, Abirami S, and Trueman TE. Multilabel Aspect-Based Sentiment Classification for Abilify Drug User Review. *In 2019 11th International Conference on Advanced Computing (ICoAC)*, IEEE, 376-380, 2019. https://doi.org/10.1109/ICoAC48765.2019.246871

[21] Baadel S, Thabtah F, Lu J, and Harguem S. OMCOKE: A Machine Learning Outlier-based Overlapping Clustering Technique for Multi-Label Data Analysis. *Informatica*, 46(4), 2022. https://doi.org/10.31449/inf.v46i4.3476

[22] Yoon Kim. Convolutional neural networks for sentence classification. 2014. https://arxiv.org/abs/1510.03820

[23] Oo SH, Hung ND, and Theeramunkong T. Justifying convolutional neural network with argumentation for explainability. *Informatica*, 46(9), 2023. https://doi.org/10.31449/inf.v46i9.4359

[24] Burkhardt S and Kramer S. Online multi-label dependency topic models for text classification. *Machine Learning*, 107(5):859-886, 2018. https://doi.org/10.1007/s10994-017-5689-6

Augmented Reality in Sports Education and Training for Children With an Autism Spectrum Disorder

Adel Fridhi, naila Bali, zied Hassen
Research Laboratory on Disability and Social Unsuitability, LR13AS01, ISES, UMA, Tunisia
Higher Institute of Special Education, Tunisia
E-mail : adel.fridhi2013@gmail.com, naila.bali2020@gmail.com, zied.hassen2020@gmail.com

Keywords: augmented reality, autism spectrum disorders, adapted physical activity, avatar, daily environment

Received: November 1, 2022

3D modeling and augmented reality (AR) offer innovative perspectives for training in sports activity for children with autism spectrum disorders (ASD). The objective of this article is to offer a reflection on the design and learning methodology in the field of adapted physical activities (APA), with the aim of improving its credibility towards children with ASD. We present an original experience of development by augmented reality in team sport: an ergonomic approach to activity in a natural situation makes it possible to model the decision-making of children with ASD; this is used to guide children with ASD to follow the existing avatar in the scene in a daily environment (DE) using augmented reality (AR).

Povzetek: Predstavljen je sistem za razvoj športnih sposobnosti avtističnih učencev.

1 Introduction

The term "autism spectrum disorder" (ASD) corresponds to a disorder may be mental which attracts the intense attention of psychologists, neurophysiologists.

Due to the development of new technologies the construction of developed versions of virtual reality (AR) has become possible. AR is simulated using modern computer techniques and an everyday environment (including a variety of interacting avatars), which is perceived by the subject through certain interface. The applications of AR may include entertainment (eg, adapted physical activity) and educational purposes, the scope of its application is currently being considerably broadened. AR is an interactive experience of a real environment where certain things residing in the real world (objects, people, events, etc.) are enhanced by computer-generated perceptual information. AR can be defined as a three-functional system: it combines real and virtual, it provides real-time interaction, and it enables accurate 3D recording of real and virtual objects. [1]

Direct physical activities with the teacher applied to children with ASD remain a classic method. However, the ability of such activities to replicate real-life tasks is obviously questionable. At this level, AR seems to be much more useful because it increases the capacity of memory, gaze, objective and exact control of the different behaviors of ASDs. The evaluation of the attentional, emotional and executive functions of a subject by the creation of AR has above all aroused our interest. The potential contribution of AR in APA to the assessment of children with ASD appears promising, and these children may benefit from this technological innovation.

As it was mentioned, the condition of children with ASD is usually assessed by classical methods, which are now traditional at the same time, several 3D tasks have also been digitally reproduced using new technologies. However, "Would AR be a more appropriate modality to help children with ASD perform and successfully learn physical activities?" is an unresolved question. The objective of our paper is to provide an answer to this problem by proposing a new method of learning APA using AR for this population. First, the cognitive functioning of children with ASD will be discussed according to different explanatory hypotheses, then an evaluation of the emotional and attentional advantages via the educational scenes in APA will be discussed and their use in relation to children with ASD, on the one hand, as a means of intervention for this type of population and, on the other hand, as a means of evaluation in APA. Finally, the discussion will focus on whether AR would be a more appropriate way to assess and improve the behavior of children with ASD.

The query is as follows: "How the use of AR concerning children with ASD can contribute to the improvement of the observation, the memorization of the APA realized in 3D." To address this problem, we will focus on two dimensions: the memory test and the repetition of APAs in children with ASD.

2 Increased reality in adapted physical education processes for asd

AR is a new technology that has added the most interest in the field of adapted physical education [2], as it allows real-time interaction and immersion between real and virtual space [3]. AR comprises different stages which increase the complexity of the effect they produce and require highly developed computer equipment, a marker, a camera and then virtual elements (avatars) are added to the real space [4]. This techno-pedagogical tool used in physical education allows a huge number of possibilities that promote meaningful learning for children with ASD through the approach of classroom contexts with which they will be in full immersion and interaction with avatars [5, 6].

The strength of AR's success lies in the large number of benefits it brings to physical education adapted to children with ASD [7]. From our practical tests we noticed that it considerably increased the motivation of children with ASD, because it allowed to present the content of the motor actions existing in the scene projected on the screen in an attractive way, arousing curiosity for the process of learning [8-9]. Methodological trends in physical education aim to increase motivation at the highest level [10].

Our experiments show that AR contributes to the promotion of the contents of the scene in an autonomous way and from a constructivist approach, produced by learning by discovery [11]. It also facilitates collaborative sequencing as another potential of AR is its adaptation to the different stages of adapted physical education [12]. It is a very versatile tool that can be adapted to the profiles of children with ASD allowing it to be used effectively [13]. In addition to all of these factors, AR contributes to sports education by improving digital skills and bringing virtual content into the try-out room, which allows ASDs to contextualize their new adapted physical activity learning method. and to facilitate the understanding of different stages [14,15]. The end result is a great improvement in academic performance in the showroom [16].

The physical education of ASDs is a pedagogical subject that greatly benefits from the versatility of AR, because it allows this population to increase their performance, it facilitates the understanding of the practical tests of its content [17] and accelerates the development of complex motor skills [18] and improves spatial orientation and interpersonal skills in ASD [19].

AR can be an excellent weapon for children with ASD if used to model a learning environment mixed with physical training that encompasses all knowledge in adapted physical education [20]. The great value of AR as a tool in the field of physical education adapted for ASD is becoming very evident [21].

3 Methodology

a. Memory test

We started by using AR to measure the memory of children with ASD to focus on faces and avatars. The work consisted of memorizing a number of color sets and avatars displayed virtually with a head-mounted display and they had to indicate whether they are identical or not (Fig. 01). The results showed that children with ASD are highly motivated to recognize objects. This review therefore reflects what has been concluded in the scientific literature: children with ASD memorize a smaller number of faces.

Figure 1: Presentation of color game to test memory.

Finally, AR has been used as part of the assessment or verification of ASDs due to its scientific advantages that produce faces or situations of adapted physical activity easily reflecting the real environment. The results mentioned above are consistent and very interesting with what is included in the scientific analysis regarding ASD. AR actions prove important eye contact, very easy social decision making, great respect for social conventions and great attention to memorizing faces. AR would then prove to be a valuable assessment and verification tool for ASDs because it offers to objectively monitor the behaviors associated with this diagnosis. AR technology aims to change the approach to assessment and memorization of ASD-related processes. The next phase of this progression will allow access to remote, AR-based environments (cases of covid-19 during containment). Finally, before reaching this goal, several questions must be addressed and resolved [22]. Applications modeled and created by AR are used successfully to train children with ASD to overcome memorization difficulties (when adding an avatar located in the same scene) (Fig. 02).

Figure 2: Addition of another color to test the memory of the TSA.

When we talk about augmented reality, we think of a new technology in its development and use. But as surprising as it may seem augmented reality has not been used in the setting of physical activity suitable for children with ASD, our goal is to help this population as best as possible. Augmented reality is a technology of the future, but we like to think of it as an innovation of the present. In this article we will go through some examples of the use of AR in the context of adapted physical activities more specifically towards children with ASD.

The example below (Fig. 03) is the first use of augmented reality in the context of adapted physical activities for children with ASD. AR was used using a marker located in an exhibit hall, children could see color avatars appear. When you rotate the marker you can look at the virtual object modeled in 3D (Fig. 03).

b. Repetition test for sports activities

Figure 3: First try.

Overall, this step has been very well accepted by ASDs. We noticed an improvement in the eyes, that's why we considered it a very interesting step and it can be improved in the following steps. Anyway, after the success of AR technology, this system was further improved in order to be the most stable, the most real and also to allow its use by this subject, the camera at the top. That allows coverage of the space used is visible by TSAs in the showroom (Fig. 04).

Figure 4: Second try.

These experiences have given us the ability to know memory capacity and even see real actions by ASDs. Our research laboratory Handicap and social maladjustment RLHSM (Research Laboratory on Disability and Social Unsuitability therefore released an application that allowed children with ASD to "memorize, follow the adapted physical activities and apply the existing phenomenon in the existing scene" through the RA.

4 Results

The following table (01) shows the benefits of AR in sports for children with autism spectrum disorder, and then you will see an experimental study (fig.03, 04, 05, 06).

Benefit of augmented reality	Results
Build a training environment	-The AR solution enables the creation of a personalized training environment for ASD child that displays forecast information for athletes. Users can practice playing directly with virtual opponents created by computers before participating in real matches. Players using AR sports solution can experience this type anywhere. Additionally, players can interact directly within the app. It is something special in this type of technology that virtual reality (VR) does not have.
Virtual stadium in development	Fans can participate in home game commentary. Augmented reality in sports provides extremely professional graphical analysis. AR allows them to easily connect to stadiums in the virtual world. Sports companies incorporate this model with the aim of helping users visualize ball curves/trajectories and other important live match details.
Ease of learning	- it is important first of all to master this new learning method. This being in order to be able to take full advantage of all that it has to offer in the field of sport. - Once you understand how augmented reality works, you can then learn much more easily by having a much more precise and concrete view of the subject.

Practice anytime/anywhere	AR used in sports training can help ASD child, as well as their coaches, set new skill goals, as the sports AR app will display information to users over time. Real about every shot, throw and mile traveled. The practice session will help the players to know better techniques to improve. Moreover, coaches or sports professionals can evaluate training sessions with AR to make the right decision for their athletes.
Link building	-When a fact becomes more real, links are created more simply and this is a very great advantage of augmented reality, including the world of ASD children's sports without any limits.
A more real feeling	- Augmented reality brings a much more real feeling when it is requested. For this, she uses all our senses to help us ASD children.
Major discoveries	Augmented reality sometimes gives us the ability to see what has never been seen before, especially the ocean floor. Because of all this, discoveries and new studies continue to increase every day so that we know better and better the world in which we live.

Table 1: Benefit and results of augmented reality.

AR technology has the capacity to meet all the requirements formulated [23]. Reality AR makes it possible to simulate virtual sports environments, in a very ecological way (Fig. 05), in particular inserted in a demonstrative situation (in a classroom). Augmented Reality (AR) offers the possibility of managing the control of distracting components, the complexity of stimuli and their alteration. Dynamically variable, in response to the actions of children with autism. Other technical specifications of the responses (precision, rhythm) can be collected to allow a more precise analysis.

This new methodology can increase the reliability of the assessment by reducing the variability due to the differences between the sports educators, the virtual testing environment and the efficiency of the modeled avatars. Finally, it can improve its validity by allowing more detailed and specialized behavioral investigations and by increasing the ecological characteristics of what is measured with children with ASD [04].

A large number of experiments were done for the present investigation and a group of ASDs was also made available for the development of the experiment (Fig. 05 & 06). A new innovative methodology, since the process of education was applied to it through the use of AR. The objective was to find out if there were other significant ones in each of the dimensions in the experimental group. That is to say whether the method applied in one group and another influenced each of the dimensions of the study.

Figure 5 : Adaptation of TSA children with the scenes presented.

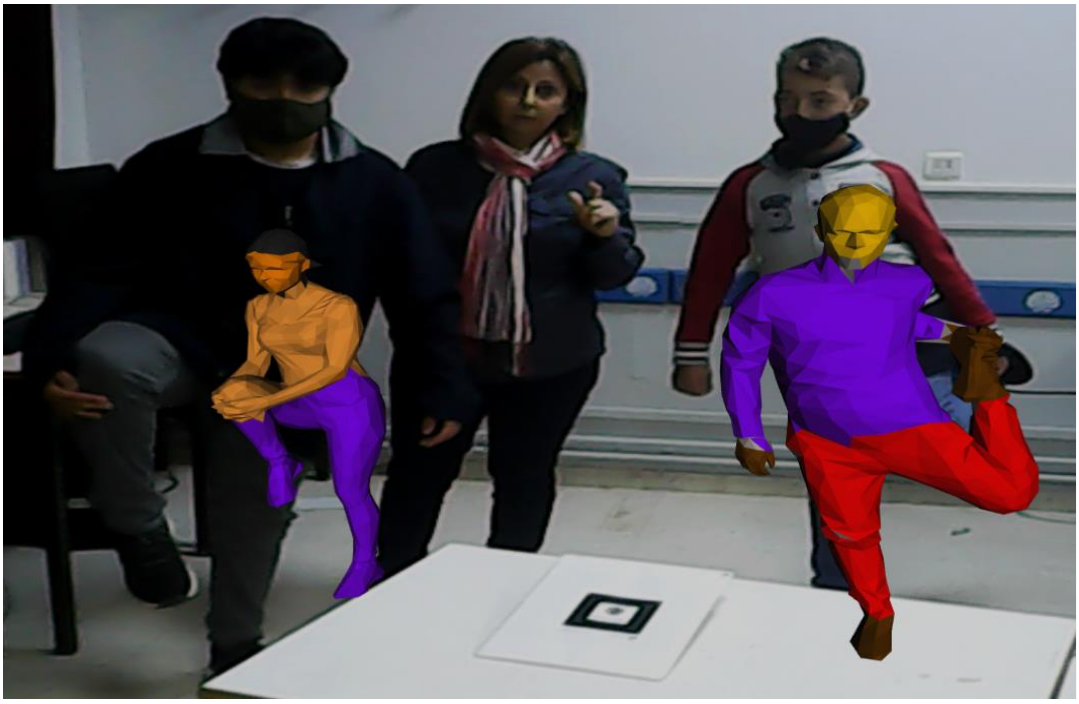


Figure 6 : Confirmation of the adaptation of children with ASD with the scenes presented.

5 Discussion

The integration of new technologies in exhibition spaces is becoming an increasingly common practice, more precisely, in the field of physical education, more and more training is betting on these tools in the exercise of their practice. Educational, because of the benefits and emotional enhancement they bring to children with ASD [25]. As a result of this new learning method, the present study aimed to analyze from a practical perspective how the application of AR affected huge several skill variables in ASD. The results obtained in the analyzes after carrying out the experiment show how the experiment with AR improved all of the variables evaluated. We continue on a line of research that consolidates the idea of the increased motivation caused by practical trials with the use of this new technology, [26]. Also, making contact with this technology allowed ASDs to improve their perception of physical education and aroused their interests and knowledge.
Finally, and on the basis of the few questions asked at the beginning of this paper, the results obtained provided motivating answers to go very far with this new learning method in adapted physical education.

6 Conclusion

The techniques of AR are constantly changing the approach to the assessment and rehabilitation of ASD-related processes. The next phase of this progress will allow one to provide access via the Internet to environments based on virtual or augmented reality. However, before this vision is realized, several questions will have to be addressed and resolved [27]. Applications created by AR are used to train children with learning difficulties, such as children with ASD, on certain safety gestures.
This paper is a review, and thus the correspondence of the study to the existing ethical standards does not need special confirmation. At the same time, some studies of the authors of the review have been mentioned in the latter; in all these studies involving human subjects, the existing international ethical norms have been strictly observed [28].

7 References

[1] Fridhi, A., Bali, N., Rebai, N., *et al.* Geospatial Virtual/Augmented Environment: Applications for Children with Pervasive Developmental Disorders. *Neurophysiology*, 2020, p. 1-8.

[2] López-Faican, L.; Jaén, J. EmoFindAR: Evaluation of a mobile multiplayer augmented reality game for primary school children. Comput. Educ. 2020, 149, 103814.

[3] Madanipour, P.; Cohrssen, C. Augmented reality as a form of digital technology in early childhood education. Australas. J. Early Child. 2020, 45, 5–13.

[4] Lester, S.; Hofmann, J. Some pedagogical observations on using augmented reality in a

vocational practicum. Br. J. Educ. Technol. 2020, 51, 607–866.

[5] Lee, I.J. Kinect-for-windows with augmented reality in an interactive roleplay system for children with an autism spectrum disorder. Interact. Learn. Environ. 2020, 1–17.

[6] El Kabtane, H.; El Adanani, M.; Sadgal, M.; Mourdi, Y. Virtual reality and augmented reality at the service of increasing interactivity in MOOCs. Educ. Inf. Technol. 2020, 1–27.

[7] Salar, R.; Arici, F.; Caliklar, S.; Yilmaz, R.M. A Model for Augmented Reality Immersion Experiences of University Students Studying in Science Education. J. Sci. Educ. Technol. 2020, 1–15.

[8] Rivadulla, J.C.; Rodríguez, M. Incorporation of augmented reality in science classroom. Contextos educativos. Rev. Educ. 2020, 25, 237–255.

[9] Cabero, J.; Roig, R. The motivation of technological scenarios in augmented reality (AR): Results of di_erent experiments. Appl. Sci. 2019, 9, 2907

[10] Demitriadou, E.; Stavroulia, K.E.; Lanitis, A. Comparative evaluation of virtual and augmented reality for teaching mathematics in primary education. Educ. Inf. Technol. 2020, 25, 381–401.

[11] De Almeida, G.N.; Cabero, J. Aid-augmented reality for reinforced concrete class: Students' perception. Alteridad 2020, 15, 12–24

[12] Rodríguez, A.M. ; Hinojo, F.J. ; Ágreda, M. Diseño e implementación de una experiencia para trabajar la interculturalidad en Educación Infantil a través de realidad aumentada y códigos QR. Educar 2019, 55, 59–77.

[13] Sahin, D.; Yilmaz, R.M. The e_ect of Augmented Reality Technology on middle school students' achievements and attitudes towards science education. Comput. Educ. 2020, 144, 103710.

[14] Habig, S. Who can benefit from augmented reality in chemistry? Sex di_erences in solving stereochemistry problems using augmented reality. Br. J. Educ. Technol. 2019, 51, 629–644.

[15] Arici, F.; Yildirim, P.; Caliklar, S.; Yilmaz, R.M. Research trends in the use of augmented reality in science education: Content and bibliometric mapping analysis. Comput. Educ. 2019, 142, 103647.

[16] Fidan, M.; Tuncel, M. Integrating augmented reality into problem-based learning: The e_ects on learning achievement and attitude in physics education. Comput. Educ. 2019, 142, 103635.

[17] Hsiao, K.F. Using augmented reality for student's health—Case of combining educational learning with standard fitness. Multimed. Tools Appl. 2013, 64, 407–421.

[18] Chang, K.E.; Zhang, J.; Huang, Y.S.; Liu, T.C.; Sung, Y.T. Applying augmented reality in physical education on motor skills learning. Interact. Learn. Environ. 2019, 1–13.

[19] Gallego-Lema, V.; Muñoz-Cristobal, J.A.; Arribas-Cubero, H.F.; Rubia-Avi, B. Orienteering in the natural environment: Ubiquitous learning through the use of technology. Movimento 2017, 23, 755–770.

[20] Aznar-Díaz, I.; Cáceres-Reche, M.P.; Trujillo-Torres, J.M.; Romero-Rodríguez, J.M. Mobile learning y tecnologíasmóviles emergentes en Educación Infantil: Percepciones de los maestros en formación. Rev. Espac. 2019, 40, 14–21.

[21] Moreno-Guerrero, A.J. ; Rodríguez-Jiménez, C. ; Ramos, M. ; Sola-Reche, J.M. Interés y Motivación del Estudiantado de Educación Secundaria en el uso de Aurasma en el Aula de Educación Física. Retos 2020, 38, 333–340.

[22] Fridhi, A., & Bali, N. (2021). Science Education and Augmented Reality: Interaction of students with Avatars Modeled in Augmented Reality. *International Journal of Environmental Science*, *6*.

[23] A. A. Rizzo, M. T. Schultheis, K. A. Kerns, and C. Mateer, "Analysis of assets for virtual reality applications in neuropsychology," *Neuropsychol. Rehab*., 14, 207-239 (2004).

[24] Fridhi, A., Benzarti, F., Frihida, A., & Amiri, H. (2018). Application of Virtual Reality and Augmented Reality in Psychiatry and Neuropsychology, in Particular in the Case of Autistic Spectrum Disorder (ASD). *Neurophysiology*, *50*(3), 222-228.

[25] Aznar-Díaz, I. ; Trujillo-Torres, J.M. ; Romero-Rodríguez, J.M. Estudio bibliométrico sobre la realidad virtual aplicada a la neurorrehabilitación y su influencia en la literatura científica. Revista Cubana De Información En Ciencias De La Salud 2018, 29, 1–11.

[26] Gómez-García, G. ; Rodríguez-Jiménez, C. ; Marín-Marín, J.A. La trascendencia de la Realidad Aumentada en la motivación estudiantil. Una revisión sistemática y meta-análisis. Alteridad 2020, 15, 36–46.

[27] Fridhi, A., Benzarti, F., Frihida, A., & Amiri, H. (2018). Application of virtual reality and augmented reality in psychiatry and neuropsychology, in particular in the case of autistic spectrum disorder (ASD). *Neurophysiology*, *50*(3), 222-228.

[28] Fridhi, A., Benzarti, F., Frihida, A., & Amiri, H. (2018). Application of virtual reality and augmented reality in psychiatry and neuropsychology, in particular in the case of autistic spectrum disorder (ASD). *Neurophysiology*, *50*(3), 222-228.

A Novel Method for Human MRI Based Pancreatic Cancer Prediction Using Integration of Harris Hawks Varients & VGG16: A Deep Learning Approach

Rama Prakasha Reddy Chegireddy1*, A Sri Nagesh 2
[1]Research Scholar, Department of Computer Science and Engineering, D.r. Y.S.R ANJU College of Engineering and Technology, Acharya Nagarjuna University, Guntur-Andhra Pradesh, India.
[2]Professor, Department of Computer Science and Engineering, RVR&JC College of Engineering, Guntur, Andhra Pradesh, India.
E-mail: reddysinfo@gmail.com, asrinagesh@gmail.com
* Corresponding author

Received: October 4, 2022

Keywords: BADF, classification, CLAHE, deep learning, pancreatic cancer, segmentation, UNET, medical image processing, image segmentation

Among all cancers, pancreatic cancer has a very poor prognosis. Early diagnosis, as well as successful treatment, are difficult to achieve. As the death rate is increasing at a rapid rate (47,050 out of 57650 cases), it is of utmost importance for medical experts to diagnose PC at earlier stages. The application of Deep Learning (DL) techniques in the medical field has revolutionized so much in this era of technological advancement. An analysis of clinical proteomic tumor data provided by the Clinical Proteome Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma (CPTAC-PDA) at the National Cancer Institute was used to demonstrate an innovative deep learning approach in this study. This includes a) collection of data b) preprocessed using CLAHE and BADF techniques for noise removal and image enhancement, c) segmentation using UNet++ for segmenting regions of interest of cancer. Followed by, d) feature extraction using HHO based on CNN and e) feature selection using HHO based on BOVW for extracting and selecting features from the images. Finally, these are subject to the f) classification stage for better analysis using the VGG16 network. Experimental results are carried out using various state-of-art models over various measures in which the proposed model outperforms with better accuracy:0.96, sensitivity:0.97, specificity:0.98, and detection rate:0.95.

Povzetek: Opisana je metoda globokega učenja za napovedovanje raka na ledvicah.

1 Introduction

The death rate from pancreatic cancer (PC) in the United States is among the highest of all cancers. Despite aggressive treatment approaches and combination modalities, the 5-year survival rate remains 5%. According to 2017's SEER data [1], Pancreatic ductal adenocarcinoma accounted for 47,050 deaths and new cases of 57,600 were reported. In 2030, PDAC is expected to overtake cancer as the 2nd largest cause of mortality [2]. Only 15 to 20% of sufferers are qualified for a potentially curative surgery because of non-specific indications and late discovery [3]. Whipple surgery left pancreatectomy and complete pancreatectomy+ are the three surgical options for pancreatic cancer treatment. By analyzing the resection tissues, it will be possible to determine whether or not lymph nodes are metastasizing from the tumor, as well as whether there is pre-invasive pancreatic intraepithelial neoplasia. Further therapeutic management will be based on pathological results [4]. It is important to identify neoplastic cells from benign or inflammatory cells to have a clear picture of the tumor. Because of the tremendous heterogeneity between and within tumors in growth pattern, cytology, and stroma (figure 1), this can be a daunting task. A fibrotic and inflammatory microenvironment contributes to the heterogeneity and complex growth pattern of tumors, with the latter constituting most of the tumor mass [5]. On microscopic examination, PDAC is primarily glandular, with extensive desmoplastic stroma formation. However, other structures can also be observed, including (micro-)papillae, solid nests, cribriform, or small, single-cell tumors [6]. There are several molecular factors associated with the development of non-glandular, histologically poorly differentiated tumor growth patterns, such as mesenchymal phenotypes, proteases, and neutrophil infiltrates [7,8]. PDAC grows in a dispersed pattern. It is in these cases that the tumor cells are not usually grouped, but are instead found in cellular clusters which encroach on the surrounding tissues, nerve

sheaths, and vascular networks [9]. A PanIN (Pancreatic Intraepithelial Neoplasia) is the precursor lesion of PDAC (Figure 2), and it is analogous to ductal epithelial carcinomas in colon cancers, in which ductal cells proliferate to become cancer-invasive.

Figure 1: Pancreatic Cancer: MRI image of risk patients.

A healthy pancreas and chronic pancreatitis have glandular and ductal features grouped in an organoid-lobular configuration, while a malignant pancreas has tumor glands that are dispersed throughout the stroma, distorted, and display solitary cells [10,11]. Chronic pancreatitis is characterized by fibrosis, ductal tissue loss, and acinar thinning, all of which were linked to an increased risk of invasive carcinoma [12]. PDAC review time for slides with histological microarchitecture, distributed development, varied microclimates, preinvasive lesions, inflammatory tissue, and sealed anatomical tissue is predicted to be 1 to 2 minutes per slide [13]. The time variable is significant for diagnosing, even if the accuracy of diagnosis is high, and it will become even more significant as the overall number of specialist pathologists' declines, and as the general demand for information and specialization increases, as well as the number of patients [14,15].

Techniques which enable and promote morphological-based tissue slide evaluation and flag crucial regions for further study by professional pathologists are thus necessary. Digital pathology has evolved as a means for evaluating histopathology slides, supporting routine diagnostics and research, as well as ensuring quality control. Reproducible tissue categories are very important in spatial tissue studies. Deep learning methods have previously been demonstrated to be effective in determining lymph node metastases and classifying tumor subsets [16].

1.1 Research gap

By identifying the onset period, pancreatic disease could be reduced from being the leading cause of death. One of the most difficult tasks completed by the radiologist up to this point has been identifying the nodules in the stomach wall. Nodules of the pancreas have diverse shapes and sizes, which makes it difficult to identify small nodules. While segmenting the tumor region, difficulties such as over-segmentation and under-segmentation can develop. While there are many imaging modalities available, using the more reliable and convenient modality is important for early tumor detection. To identify and characterize the tumor's location, scientists have recommended some procedures. The contrast of MRI for soft tissues is better than CT, and it can differentiate fat, water, muscles, and other soft tissues more easily than CT. Additionally, MRI has a higher sensitivity (33%) for detecting tumors than CT (11%). The primary goal of this research is to suggest a better framework that will detect and classify pancreatic cancer from MRI images to support radiologists in making diagnostic decisions.

1.2 Key highlights

This article aims to optimize methods and propose a framework for detecting and classifying pancreatic cancer using deep learning and image processing techniques. The primary objectives of this article are as follows:

- •To suggest a framework based on MRI images to detect and classify pancreatic cancer.
- •To improve the MRI image quality using Boosted Anisotropic Diffusion Filter (BADF) and contrasted limited adaptive histogram equalization (CLAHE) algorithms.
- •To use the UNet++ architecture to create a Computer-Aided Detection method (CAD) for the early identification of pancreatic cancers. The pancreatic region associated with a lesion is precisely separated from the MRI image by segmentation using the UNet++.
- •To extract the best subset of texture features to enhance classification accuracy and to create a classification system based on these texture features using HHO-based CNNs and HHO-based Bags of visual terms.
- •To distinguish different levels of malignancy in an MRI image by developing a classifier based on the VGG 16 model.
- •To perform quantitative analysis for various tumor classes and the accuracy of the proposed classifier is assessed against the state-of-the-art work's performance.

Organization of the paper: As we already came across the overview of PDAC and its respective areas in Section 1, part 2 discusses the literature review, third part illustrates the overall methodology adopted. The fourth part presents the performance analysis, and the fifth section summarizes the conclusion.

2 Literature review

Tonozoko et al. (2021) [17] developed a Computer-Aided Diagnostics (CAD) approach that used deep learning assessment of EUS pictures (EUS CAD) to distinguish between persons with chronic pancreatitis and those with Pancreatic Ductal Carcinoma (PDAC).

Liu et al. (2020) [18] used a CNN to determine whether patches were carcinogenic. According to the fraction of patches designated as carcinogenic by the CNN and the trained and validation datasets, a criterion for identifying pancreatic cancer was created. Researchers utilized a localized test group (101 pancreatic cancer patients and 88 controls, local test group 2) in addition to data from the United States (281 pancreatic cancer patients and 82 controls). In this study, EM algorithms and Gaussian Mixture models were integrated to highlight the most necessary properties of the CT scan, and threshold values were used to determine the percentage of tumors present in the pancreas.

Vaiyapuri et al. (2022) [19] introduce an intelligent deep-learning-enabled decision-making medical system for pancreatic tumor classification (IDLDMS-PTC) using CT images. The IDLDMS-PTC model derives an emperor penguin optimizer (EPO) with multilevel thresholding (EPO-MLT) technique for pancreatic tumor segmentation. A MobileNet model is applied as a feature extractor with optimal autoencoder (AE) for pancreatic tumor classification. To optimally adjust the weight and bias values of the AE technique, the multileader optimization (MLO) technique was utilized.

Abbas et al. (2021) [20] suggest a Computer Aided Diagnosis (CAD) system that uses Synergic Inception ResNet-V2, a deep convolutional neural network architecture, to identify PC cases from publicly available CT images. This system could extract PC graphical functionality to include clinical diagnosis before the pathogenic examination, freeing up valuable time for disease prevention. To demonstrate the relatively encouraging outcomes in terms of accuracy in recognizing BC-infected patients, simulation results using MATLAB are provided in the study. The suggested deep learning approach achieves an accuracy of 99.23%.

Li et al. (2022) [21] offer a deep-learning segmentation technique for pancreatic cancer based on a dual meta-learning framework. This can combine generic tumor data from idle MRIs with prominent tumor information from Ct scan images to improve the discrimination of high-level features. To provide rich intermediate explanations for a meta-learning technique that would follow, the randomized intermediary modality between CTs and MRIs was originally developed to fill in visual gaps.

Author	**Algorithm**	**Metrics**	**Strength**	**Weakness**
Tonozoko et al. (2021) [17]	AlexNet	AUROC – 0.924 Sensitivity – 90.2 Specificity – 74.9	Higher-resolution EUS images are used. Higher sensitivity.	Risks and feasibility of EUS imaging.
Fu et al. (2021) [19]	Inception V3	Accuracy - 0.953	Patch-level and WSI-level approach improves the overall classification accuracy	The algorithm recognizes cancer cells mainly from nuclear features. Hence prone to false positive results.
Liu et al. (2020) [18]	VGG-16	Sensitivity - 0·973, Specificity - 1·000, and Accuracy - 0·986	Achieved an accuracy approaching 99% and missed fewer tumors compared with that of radiologists.	Uses CT scans which show less tumor detection sensitivity of 11% compared to MRI (33%).
Abbas et al. (2021) [20]	ResNet	Accuracy - 99.23	The isolateral filter enhances the quality of poor images during preprocessing.	Uses CT scans which show less tumor detection sensitivity of 11% compared to MRI (33%).
Li et al. (2022) [21]	GoogleNet	Dice score - 64.94	Dual meta-learning framework for pancreatic cancer using MRI as well as CT. Outperforms state-of-the-art methods based on CT imaging.	NA

Table 1: Summary of literature review.

3 Methodology

This section outlines a novel approach for classifying pancreatic cancers based on the Pancreatic Ductal Adenocarcinoma cohort of the Clinical Proteomic Tumor Analysis Consortium (CPTAC-PDA) dataset. While there are various imaging techniques available, MRI demonstrates improved tumor detection sensitivity, which aids in discovering smaller tumors (Grade I). The novelty of this study is the application of image-enhancing methods and optimization strategies to MRI images to increase the classification accuracy when compared to the state-of-the-art research under discussion. The overall design of the proposed framework is shown in Figure 3, with the steps outlined below.

During the pre-processing step, CLAHE and BADF are used to enhance the images obtained from the publically available MRI image collection CPTAC-PDA. A source image is divided into non-overlapping contextual components known as sub-images, tiles, or blocks by the CLAHE method. To balance each contextual area, the CLAHE approach uses histogram equalization. The cropped pixels are then redistributed throughout the grey levels after the original histogram is cropped. While traditional histograms, redistributing histograms cap pixel intensities at a maximum value. By including a Partial Differential Equation (PDE) after it generates the diffuse image, the suggested BADF improves on the existing anisotropic diffusion filter. It's a sophisticated unsupervised machine learning-based image enhancement tool. It's also feasible to smooth details with a diffusion process that's weak at the edges and borders of the images and not only smooths out the image but also preserves important characteristics like edges and patterns. Excellent results were achieved when the number of iterations was set to 20 based on extensive testing. Once images are preprocessed, Segmentation is carried out which is a crucial part of an image classification method where the MRI image is segmented to isolate the nodules. In this work, the UNet++ architecture is used for the segmentation of MRI images. Once segmented regions are obtained, features are extracted and selected by using HHO-based CNN and HHO-based BOVW. After segmentation and feature extraction, the segmented tumor is identified using texture features. Finally, the VGG-16 model is used to distinguish between normal and tumor grades from the MRI images. The Convolution Neural Network (CNN) architecture VGG-16 is one of the best models for image classification which allows transfer learning. Transfer learning is the process of applying the knowledge gained from one problem to another related problem for further improvement.

3.1 Data collection

A dataset of CPTAC-PDA pancreatic ductal adenocarcinomas from the National Cancer Institute is included here. Proteogenomic, a large-scale method of studying cancer genetics, is the goal of CPTAC [22].

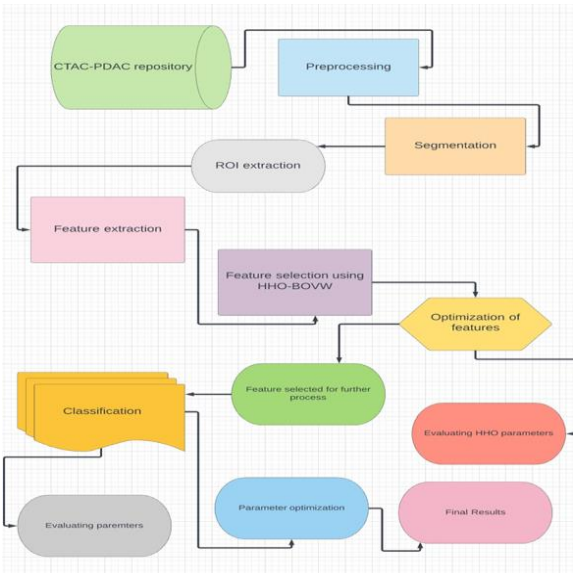


Figure 2: The overall architecture of the proposed framework.

The Cancer Imaging Archives is collecting radiology and pathology images from CPTAC patients to provide researchers with access to these images so they can investigate cancer phenotypes and correlate them with proteomic, genomic, and clinical findings.

There is a TCIA Collection for each type of cancer, called CTAC- cancer type, which stores the images for each type. Radiology pictures are compiled from routine imaging conducted on patients immediately before pathology diagnoses, as well as follow-up scans where available. As a result, in terms of scanner modalities, vendors, and acquisition processes, radiology picture information sets are varied. The CPTAC (Figure 4) qualification method includes collecting pathology images. The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma (CPTAC-PDA)1 contains 45786 pancreatic images from CPTAC third-phase patients. A total of 45 radiology topics and 77 pathology topics [23] are included. This dataset includes samples from CT, CR, and MRI scans. The pictures are of various sizes, but they were shrunk to 128 in the current work. The flexibility of the answer produced by the diverse qualities of various imaging techniques is increased by using numerous modalities in the training step.

3.2 Preprocessing

Preprocessing is carried out for removing noise and anomalies and also thereby enhancing the images for better prediction. So here we use both CLAHE and BADF. They will be compared in Table 1 over measures like PSNR, SSIM, and MSE for better preprocessing analysis over ADF, BADF, AHE, and CLAHE. From this, we get to know that, the higher the PSNR and SSIM, the lower the MSE will give many accurate results.

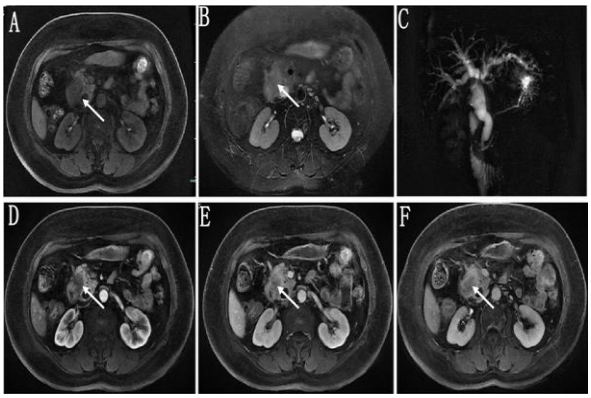


Figure 4. Pathology confirmed pancreatic ductal adenocarcinoma in an elderly female patient. On fat-suppressed LAVA T1-(A) and T2-(B) weighted imaging, C) MRI Cholangio-Pancreatography (MRCP), D) Gadolinium-enhanced images in arterial, E) Portal, F) dela

3.2.1 CLAHE

Because the pancreas is related to other organs such as the duodenum and gallbladder, the input volume was enhanced to make the pancreas more visible. To begin, we modified the MRI images by adding a window center (60) and window width (400) to make the abdomen visible. By boosting the contrast of the pancreatic region, the basic dataset was constructed by contrast-limited adaptive histogram equalization (CLAHE) [24-27]. By using the dynamic histogram equalization method, each pixel is mapped to its grayscale neighbors. Because the number of times the approach is used is equivalent to the number of pixels in the area, it consumes a lot of processing resources. CLAHE accomplishes this by establishing a criterion. If part of the picture's grey levels surpasses the threshold, the surplus is dispersed equally among all grey levels. The image will not be over-enhanced as a result of this processing, and the issue of noise amplification will be minimized.

3.2.2 BADF

The Perona-Malik Diffusion Process is another name for the anisotropic diffusion filter, and it is named after the people who devised it. It focuses primarily on eliminating noise while maintaining fine features in the image. In general, the filters employ the very same methodology as edge detection. Using multiple blurred pictures generated by the diffusion process, the anisotropic diffusion filtering process may be described. The proposed BADF improves on the previous anisotropic diffusion filter by adding a Partial Differential Equation (PDE) after creating the diffused image. Diffusion, which is absent at the edges and boundary, can be utilized to smooth the surface [28]. After that, four conduction operators obtained from Equations (20) and (21) are used to attenuate the high-frequency elements in each direction.

$$g_N=\frac{1}{1+(\frac{\nabla_N I_{i,j}}{k})^2} \quad (1)$$

$$g_S=\frac{1}{1+(\frac{\nabla_S I_{i,j}}{k})^2} \quad (2)$$

$$g_E=\frac{1}{1+(\frac{\nabla_E I_{i,j}}{k})^2} \quad (3)$$

$$g_W=\frac{1}{1+(\frac{\nabla_W I_{i,j}}{k})^2} \quad (4)$$

Algorithm 1: CLAHE

Inputs: I-mammogram image, seed point
Output: E- Cartesian enhanced image, P-Polar enhanced image

$I_{CL} \leftarrow$ CLAHE(I)
$I_2 \leftarrow$ NI(I), $\mu \leftarrow$ mean of all pixels in I_2
$I_3 \leftarrow$ NI(I_{CL}) .* (1-exp(-(I_2/μ)))
$I_3 \leftarrow$ NI(I_3)
$M_2 \leftarrow$ RM(I_3,[16 16])
$I_4 \leftarrow$ NHEQ(I_3)
$I_5 \leftarrow$ NI(I_3 + RM(M_2,[16 16]))
$I_6 \leftarrow$ NHEQ(I_5) .* I_4
$E \leftarrow$ NI(I_6)
$P \leftarrow$ Polar(E) with center at seed point
$\mu(r) \leftarrow$ mean{P(r,θ)} for all r; $\mu_{MAX} \leftarrow \max(\mu)$
$\sigma \leftarrow$ smallest radius r with $\mu(r) \geq 0.4\mu_{MAX}$

$$G(r,\theta) \leftarrow \begin{cases} 1 & r \le \sigma \\ \exp\left(-0.5\left[(r-\sigma)/\sigma\right]^2\right) & r > \sigma \end{cases}$$

$P(r,\theta) \leftarrow P(r,\theta)$.* $G(r,\theta)$ for all r,θ

$$P \leftarrow \left(P - mean(P)\right)/std(P)$$

```
function N = NI(X) % Normalize image
 N ← X - min(X)
 N ← N / max(N)

function N = NHEQ(X) % Normalize and hist. eq.
 N ← HistogramEqualize(NI(X))

% Function to compute the regional mean
function N = RM(X,[b1,b2])
 M ← image where each pixel is mean value
     of a b1 by b2 block B in image X
 N ← M resized to size of X using 2D bilinear
     interpolation
```

K is a scalar that controls the level of smoothness, but it must satisfy (K > 1), because a higher value of K results in smoother outcomes. In a standard anisotropi diffusion filter, K is set to 7. Equation (24) [29] is used to automatically calculate variable K based on local statistics in this investigation.

$$k=2*\left|\frac{mean(f_{i,j})}{(0.75*\sigma(f_{i,j}))}\right| \quad (5)$$

Here, the standard deviation is denoted by σ. Using Equation (10), we can determine the variance by smoothing the visuals.

$$I_{i,j} = I_{i,j} + 0.25[(g_N * \nabla_N I_{i,j}) + (g_S * \nabla_S I_{i,j}) + (g_E * \nabla_E I_{i,j}) + (g_W * \nabla_W I_{i,j})] \quad (6)$$

where Ii, j is a smoothened image.

Algorithm 2: BADF

Step 1: Double the size of the input image.
Step 2: Diff im is a PDE (partial differential equation) that needs to be initialized.
Step 3: Set the pixel distances in the centre.
dx = 1;
dy = 1;
Step 3: Identify four different 2D convolution masks (N,S,E,W).
hN = [0 1 0; 0 -1 0; 0 0 0]
hS = [0 0 0; 0 -1 0; 0 1 0];
hE = [0 0 0; 0 -1 1; 0 0 0];
hW = [0 0 0; 1 -1 0; 0 0 0];
Step 4: Before evaluating the diffusion function, identify the finite difference.

Table 2. Overall analysis under PSNR, MSE and SSIM.

Preprocessing models	**PSNR**	**SSIM**	**MSE**	**Image**
AHE	23.56	0.24	3.29	
ADF	22.8	0.33	4.71	Image 1
CLAHE	43.9	0.72	8.44	
BADF	46.2	0.85	9.31	
AHE	23.59	0.245	3.3	
ADF	22.83	0.333	4.75	
CLAHE	43.95	0.726	8.49	Image 2
BADF	46.27	0.851	9.36	
AHE	23.6	0.25	3.32	
ADF	22.9	0.34	4.79	Image 3
CLAHE	44.1	0.73	8.5	
BADF	46.3	0.86	9.4	
AHE	23.62	0.257	3.37	
ADF	22.93	0.345	4.8	Image 4
CLAHE	44.12	0.738	8.53	
BADF	46.36	0.862	9.46	
AHE	23.68	0.26	3.4	
ADF	23.2	0.35	4.83	Image 5
CLAHE	44.18	0.74	8.6	
BADF	46.4	0.87	9.5	

3.3 Segmentation

The proposed design is depicted in Figure 5a from a high-level perspective. Unet++ is based on an encoder subnetwork, which will be followed by a decoding subnetwork. Therefore, skip paths (represented in green and blue) connecting the two subnetworks have been reconstructed, and deep supervision distinguishes UNet++ from U-Net [30,31]. This is shown in red.

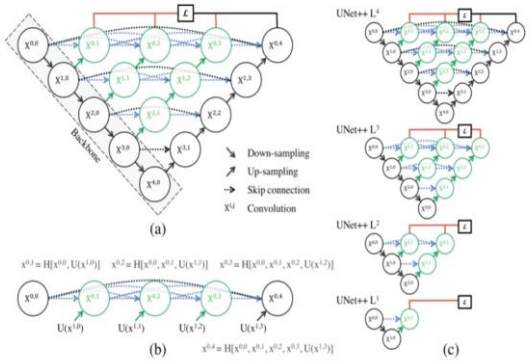


Figure 5: (a) An encoder and a decoder are linked via thick convolutional blocks in UNet++. Before fusion, UNet++ was primarily focused on bridging the semantic gap between encoders and decoders. On the original U-Net are black blocks with thick convolution blocks on skip routes in green and blue, and red deep supervision blocks. (b) A thorough investigation of UNet++'s first skip path. (c) If UNet++ was trained with a lot of supervision, it can be pruned during inference. (Color image from the internet) [33]

3.3.1 Redesigned skip pathways

The communication between the encoder and decoder sub-networks has improved thanks to redesigned skip paths. The retrieved attributes from the encoder enhance gain in the decoder in U-Net; The UNet++ method, however, uses dense convolution blocks, whose number is determined by the pyramid level.

Convolution blocks X0, 0, and X1,3, for example, contain three convolution layers. Because it is concatenated, the result of each convolution layer is merged with the reduced dense block result. Through deep convolution, features extracted from the encoder are transformed into feature maps that the decoder can decode. The ideal is considered to have a simpler approach to achieving optimum control issues if the input encoder extracting properties and accompanying decoder feature maps are conceptually equivalent.

A summary of the skip path is as follows: Let's call the result of node Xi, j xi, j while I is the encoder's down-sampling value and j is the dense block's convolution layer. The following is how to calculate the stack of extracted features denoted by xi,j:

$$x^{i,j} = \begin{cases} H(x^{i-1,j}) & j = 0 \\ H\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, u(x^{i+1,j-1})\right]\right), j > 0 \end{cases} \quad (7)$$

A convolution with an activation function of H(.) and an upsampling layer of U(.). When a node at a level j > 1 is selected, it accepts j + 1 inputs, j inputs representing previous delete paths, and lastly, its output represents the upsampled results of the lesser skip paths. Level j = 0 accepts only input from an encoder layer above it; level j = 1 accepts input from an encoder sub-network at a different stage, and level j > 1 accepts input from a lesser encoder sub-network. Because each skipping route employs a thick convolution block, all previously extracted characteristics blend and reaches the current node. Figure 5b illustrates how the characteristic mappings flow through UNet++'s top skip pathway, which better clarifies Eq. 1.

3.3.2 Deep supervision

Deep supervision is provided by UNet++ [30,31] so that the model can run in two modes: (1) accurate mode, where the categorized branches are averaged, and (2) fast mode, where one of the classification branches can be used as the categorization process map, depending on the amount of pruning in the model and the increase in the speed. In rapid mode, selecting a segment branch gives designs of variable complexity, as seen in Figure 5c.

With UNet++ one can stack skipping paths with full-resolution attributes on multiple semantic levels, including x0, j, j1, 2, 3, and 4 while being deeply supervised. Each semantic phase is assigned a loss function based on binary cross-entropy and dice coefficient:

$$L\left(Y, \hat{Y}\right) = -\frac{1}{N}\sum_{b=1}^{N}\left(\frac{1}{2}.Y_b.log\hat{Y}_b + \frac{2.Y_b.\hat{Y}_b}{Y_b+\hat{Y}_b}\right) \quad (8)$$

N is the batch size, and $\hat{Y}_b$ and Yb is the flattened projected probability and ground truth of the bth image, respectively. The difference between UNet++ and U-Net is shown in Figure 5a which includes: In terms of jump routes, (1) Convolution layers (green) improve gradient flow; (2) Closely packed skip connections on delete routes (blue); and (3) Deep supervision (red) which prevents pruning and, in the worst-case scenario, is similar to the performance of using one loss layer in model 3.3.

3.4 Feature extraction

The HHO algorithm, a new metaheuristic stochastic approach proposed by Harris hawks' behaviors, is a mathematical proposal. Harris hawks' behavior is defined by their ability to track, encircle, and approach potential prey (usually rabbits) and then attack them with excellent synchronization. Surprise pounce is a smart escape technique used in hunting. The HHO technique, like earlier meta-heuristic algorithms [34, 35], includes exploratory and exploitative steps. During the exploration phase, Harris hawks will pursue prey randomly, according to the equation:

$$X(t+1)= \begin{cases} X_{rand}(t) - r_1|X_{rand}(t) - 2_{r2}X(t)|q \geq 0.5 \\ (X_{rabbit}(t) - X_m(t)) - r_3(LB + \\ r_4(UB - LB))q < 0.5 \end{cases} \quad (9)$$

The hawks are placed at X(t + 1), the rabbit (victim) at Xrabbit (t), r1 to r4, and q are sequentially labeled from 0 to 1, Xrand (t) signifies a random selection hawk at a random location, and X_m denotes the current hawk population's average location, as computed by Equation (29):

$$X_m(t)=\frac{1}{N}\sum_{i=1}^{N} X_i(t) \quad (10)$$

Xi(t) is the place of each hawk in iteration t, and N is the total number of hawks. When the knowledge step is finished, a duration occurs between the discovery and exploitation periods. The rabbit's energy should be shaped according to Equation (30) throughout this moment of transition:

$$E=2E_0\left(1-\frac{t}{T}\right) \quad (11)$$

where E represents the rabbit's escaping energy, E_0 represents its initial energy state, and T represents the maximum number of iterations. According to the victim's physical condition, the E_0 number could vary from -1 to 1. When E0 approaches -1, the patient loses energy and vice versa. The Harris hawks suddenly approach their victim during the last stages of the algorithm's processing. There are four attack strategies available. r is a probability of escaping in this case. Harris's hawks use a delicate besiege strategy to slowly encircle the target when E ≥ 0.5 and r ≥ 0.5. The model for mathematical analysis is as follows:

$$X_i^{t+1} = \Delta X_i^t - E\left|JX_{prey} - X_i^t\right|, \Delta X_i^t = X_{prey} - X_i^t \quad (12)$$

J represents the strength of the prey's bouncing during the escape, which takes a random value between 0 and 2, and individuals present in the presence of prey are separated by a distance of Xi(t + I). The prey can't escape when E 0.5, r 0.5, due to insufficient escaping energy, and the Harris hawks' location is written as:

$$X_i^{t+1} = X_{prey} - E|\Delta X_i^t| \quad (13)$$

That is E ≥ 0.5, r < 0.5 when Harris hawks soft besiege with escalating quick dive tactics to confuse prey when the prey has the necessary power to effectively flee. It can be expressed in the following way:

$$X_i^{t+1} = \begin{cases} Y = X_{prey} E\left|JX_{prey} - X_i^t\right|, if\ f(Y) < f(X_i^t) \\ Z = Y + S \times Levy(d), if(Z) < f(X_i^t) \end{cases} \quad (14)$$

S is a 1 D random vector, where d is the problem dimension. When E < 0.5, r < 0.5, the prey has insufficient escape energy, according to the Lévy Flight function. This prey will be attacked by the Harris hawks in the following ways:

$$X_i^{t+1} = \begin{cases} X_{prey} - E\left|JX_{prey} - X_m^t\right|, if\ f(Y) < f(X_i^t) \\ Z = Y + S \times Levy(d), if(Z) < f(X_i^t) \end{cases} \quad (15)$$

After using HHO (Figure 6) for extraction, CNN is added at the end. We believe that the huge original picture lxh is specified as x in the convolutional layer. We begin by training sparse coding to extract the tiny size image from the giant picture. It is necessary to compute the f=(wxs+b) property by computing the activation function and the weights and variances between the explicit and visual layer units. We acquire the matching value f' = (wxs'+ b') for each small picture, as well as the convolution values of these fs' and the matrix of convolution of the properties, for each small image. These qualities must next be categorized after they have been obtained by convolution.

3.5 Feature selection

In four steps, the BoW model is explained. To begin, each image of the given image collection is sampled for patches represented by local descriptors. Second, a clustering algorithm generates a visual vocabulary, with each cluster center corresponding to a visual word. Third, a new image's local characteristics can be quantified using the visual vocabulary gathered earlier. Lastly, a BoW histogram is produced for image representation [36,37,38] by collecting the frequency of each bag of visuals in the frame.

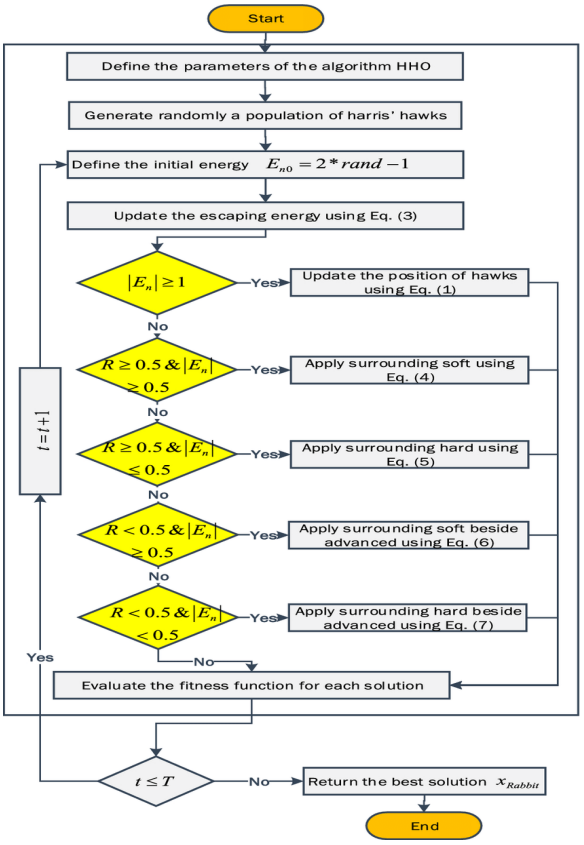


Figure 6: HHO-based flowchart for feature extraction.

As explained in the image, a set of elements from each pixel is moved to a fresh feature space with k characteristics, where k is the number of k-means centroids. Hard-assignment coding was employed to encode the features in this study. The following is an example of a BoW image representation: Provided the visual words BoW in a vocabulary,

$$X(W_i) = \frac{1}{n}\sum_{c=1}^{n} \begin{cases} 1\ if\ i = arg_j min||W_j - P_c|| \\ 0 \qquad\qquad otherwise, \end{cases} \quad (16)$$

n stands for the number of patches in the image, while pc stands for patch c. Following that, a pictorial representation is constructed using the BoW paradigm and viewed as a "bag" of visual words.

At the start, strength profiles are employed to collect the tumor's and the surrounding area's intensity difference. An intensity profile is a vector of picture intensity values calculated by analyzing the brightness of pixels along the cancer border. The pixels were taken from the center of the tumor to the border of the

cancerous area. As seen below, the intensity profile is created. Gaussian kernels smooth the spots at the tumor border to prevent them from being affected by noise, which may cause the boundary normal to shift. The Gaussian kernel is explained as follows in one dimension:

$$G_{1D}(X;\sigma) = \frac{1}{\sqrt{2\Pi\sigma}} e^{-X^2/2\sigma^2} \qquad (17)$$

To convex with the points on the cancer boundary, the first derivative of GiD (X,) is employed. The standard deviation is σ . In a picture, L(x,y) represents the coordinates of the tumor boundary. Convolution results in the points' coordinates

$$B(X^1, Y^1) = b(x,y)^* G_{1D}(X;\sigma)' \qquad (18)$$

The border normal's angles are calculated using

$$\theta = \arctan\left(\frac{y'}{x'}\right) \qquad (19)$$

For all the locations correlated with an intensity profile, the angle θ is used as the coordinate.

$$X_i = x_i + l \times cos\theta_i,$$
$$Y_i = y_i + l \times sin\theta_i \qquad (20)$$

There is a distance l between normal and border sites along the tumor boundary. This is the distance between the location on the border and the normal sites along the tumor boundary. Therefore, the picture's parameters (x1,y1) may not be exact pixel dimensions. Pixels in the picture are located using linear interpolation. Two crucial steps in building a BoW model are patch sampling and local descriptors. To simplify the subsequent computation for each raw patch, the one-dimensional feature vector is created. SIFT descriptors, which are scale and rotation invariant, are a better alternative to raw patches. Two visual vocabularies are created using precompiled patches from cancer and cancer margin regions, accordingly. As a result of this process, the vocabularies formed grow more locally unique. Another way to put it is that visual representation based on a region-specific language is more meaningful than representation based on a universal

vocabulary that uses all of the image's data.

Patches collected in the margin zone together with the four subregions are mapped to the margin region's vocabulary to generate the image representation in the margin zone. The BoW representation for the margin region is constructed by integrating the BoW histograms for each area. If the vocabulary of the margin sector contains k1 words, the BoW description of the margin area is a vector with 5*k1 dimensions. As a result, the picture now has two BoW histograms: one for the cancer zone and one for the cancer margin. Finally, the recommended region-specific BoW characterization for the malignancy on a pancreatic cancer image is created by joining these two BoW histograms together.

3.6 Classification

CNNs are learned in a feed-forward method, with error back-propagation from the classification layer to the first convolutional layer, from the very first input layer to the final classification stage. The following is an example of a forward pass: layer l's neuron I receive input from layer l-1's neuron j:

$$ln_i^l = \sum_{j=1}^{n} W_{ij}^l x_j + b_i \qquad (21)$$

Non-linearity ReLu functions are used to calculate the output:

$$out_i^t = \max(0, ln_i^t) \qquad (22)$$

Every neuron in the convolutional and fully connected layers uses equations (2) and (3) to analyze the input and receive the output in the form of nonlinear activation. The pooling layer moves a K x K square window across the N x N feature map and calculates the highest or average value of each variable. As a result, the feature map's spatial size shrinks from N×N to $N/K \times N/K$.

Finally, each cancer type's classification probability is calculated using the Softmax function:

$$out_i^t = \frac{e^{ln_i^t}}{\sum_i e^{out_k^t}} \qquad (23)$$

The back-propagation algorithm is used to train a CNN by minimizing the following cost function regarding undetermined weights W:

$$c = -\frac{1}{m}\sum_i^m \ln\left(p(y^i|x^i)\right) \qquad (24)$$

The ith sample in the training set with the label yi is Xi, and the real categorization probability is ((yi|Xi). The mini-batch cost is used to estimate the development costs, and stochastic gradient descent is used to lower the cost function C over N mini-batches. The weights are then modified in the next iteration as follows, with Wlt denoting the weights at iteration t for convolutional layer l and C denoting the mini-batch cost:

$$\gamma^t = \gamma^{[tN/m]}$$
$$V_l^{t+1} = \mu V_l^t - \gamma^t \propto_1 \frac{\partial C}{\partial W_l}$$
$$W_l^{t+1} = W_l^t + V_l^{t+1} \qquad (25)$$

Where is the layer l learning rate, γ is the scheduled rate that decreases the initial training rate α after a certain number of epochs, and μ is the momentum that determines the effects of earlier modified weights in the most recent edition.

Every iteration of training updates the weights of the CNN layers using equation (6). There are 16 layers and 138 million weights that can be learned using the VGG16 framework. Overfitting in the training and development of such deep networks can be caused by the enormous local minima in equation (5). As a result, we needed to use the pre-trained VGG16 dataset to create the weights. For limited datasets, however, determining the right local minima for the cost function in equation (5) is particularly challenging, resulting in the overfitting of the network. In this case,

weights were pre-trained on the VGG16 model [39,40].

VGG16 was fine-tuned on the PDAC dataset after the weights were transferred. This design is discussed in Figure 7, which illustrates the VGG16's thirteen convolutional layers and three fully linked layers. If we use the layer-by-layer fine-tuning technique, adding one layer at a time will result in nineteen layers. It will be essential to use 95 VGG16 designs to fine-tune five-fold cross-validation. If the training duration for each structure is roughly thirty minutes, fine-tuning the VGG16 layer-by-layer will take more than a week. Determining the appropriate parameters for layer-wise fine-tuning will take a similar length of time. The findings were slightly improved with a layer-by-layer fine-tuning method.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 7: VGG16 network trainable parameters

Based on the pooling layers, the VGG16 architecture can be divided into six blocks. Figure 7 illustrates this approach. The block-wise layout of the VGG16 is depicted in Figure 8. The final fully connected layer of VGG16 generally consists of 1000 neurons that relate to ImageNet classes. According to the classes in the PDAC dataset, the final fully connected layer of this model is made up of three neurons.

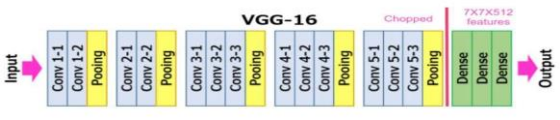


Figure 8: VGG16 architecture and its respective blocks.

4 Performance analysis

The proposed model has trained over 70% of the dataset and 30% for testing under an epoch of 10 and a learning rate of 0.09. The model is implemented using hardware specifications like Ryzen 5/7 series CPU, NV GPU, 1 TB HDD, and Windows 10 OS and software specifications like PyTorch, an open-source python library for developing deep learning models, and Google Collaboratory, an open-source Google environment for building the model. Experimental evaluation is carried over models like Alexnet, Googlenet, Inception v3, VGG19, and Resnet50 over measures like accuracy, sensitivity, specificity, recall, precision, F1-score, detection rate, TPR, FPR, and computation time. Table 2 depicts the overall analysis of various models over 5 image instances under accuracy, sensitivity, and specificity. Figure 9 depicts the graphical representation of various models over the accuracy, sensitivity, and specificity.

Table 3: Overall analysis under accuracy, sensitivity, specificity.

Models	Accuracy	Sensitivity	Specificity	Images
Alexnet	81	85	87	
Googlenet	84	89	91	
Inception v3	88	91	93	Image 1
VGG19	87	92	95	
Resnet 50	76	81	84	
VGG16	96	97	98	
Alexnet	81.3	85.4	87.1	
Googlenet	84.6	89.1	91.4	Image 2
Inception v3	88.2	91.4	93.3	
VGG19	87.4	92.5	95.2	
Resnet50	76.2	81.4	84.4	
VGG16	96.3	97.2	98.2	
Alexnet	81.5	85.7	87.3	
Googlenet	84.7	89.4	91.5	
Inception v3	88.4	91.7	93.6	Image 3
VGG19	87.6	92.7	95.4	
Resnet50	76.4	81.8	84.7	
VGG16	96.5	97.5	98.5	
Alexnet	81.8	85.8	87.6	
Googlenet	84.8	89.6	91.7	
Inception v3	88.6	91.8	93.8	Image 4
VGG19	87.7	92.9	95.6	
Resnet50	76.7	81.9	84.8	
VGG16	96.7	97.7	98.7	
Alexnet	82	86	87.8	
Googlenet	85	90	92	
Inception v3	89	92	94	Image 5
VGG19	87.9	93	95.7	
Resnet50	76.8	82	84.9	
VGG16	96.9	97.8	98.8	

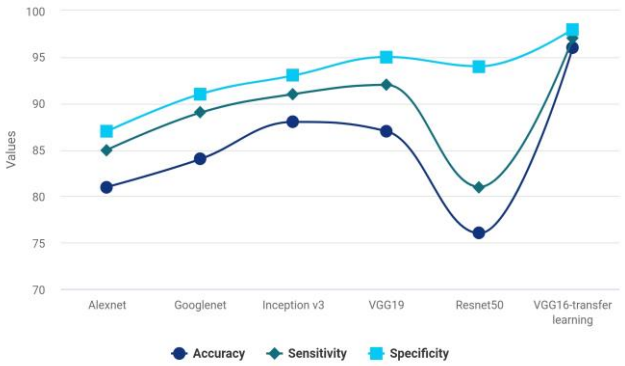


Figure 9: Models vs Measures overall analysis under accuracy, sensitivity and specificity

Table 2 depicts the overall analysis of various models under precision, recall, and F1-score. Figure 10 illustrates a graphical representation of various models.

Table 4: Overall analysis under precision, recall, F1-score.

Models	**Precision**	**Recall**	**F1-score**	**Images**
Alexnet	83	74	83	
Googlenet	82	78	86	
Inception v3	87	82	81	Image 1
VGG19	85	84	87	
Resnet 50	79	68	71	
VGG16	93	86	89	
Alexnet	83.4	74.2	83.4	
Googlenet	82.5	78.5	86.1	Image 2
Inception v3	87.2	82.1	81.4	
VGG19	85.3	84.3	87.5	
Resnet50	79.1	68.2	71.4	
VGG16	93.3	86.3	89.2	
Alexnet	83.6	74.4	83.6	
Googlenet	82.7	78.7	86.4	
Inception v3	87.5	82.6	81.6	Image 3
VGG19	85.5	84.4	87.7	
Resnet50	79.3	68.5	71.5	
VGG16	93.6	86.7	89.5	
Alexnet	83.7	74.7	83.7	
Googlenet	82.8	78.8	86.6	
Inception v3	87.8	82.8	81.7	Image 4
VGG19	85.7	84.7	87.8	
Resnet50	79.7	68.7	71.7	
VGG16	93.7	86.8	89.8	
Alexnet	84	75	84	
Googlenet	83	79	87	
Inception v3	87.9	83	82	Image 5
VGG19	86	85	88	
Resnet50	80	69	72	
VGG16	94	87	90	

Figure 10: Models vs Measures. Overall analysis under precision, recall and F1-score.

Table 3 depicts the overall analysis of various models under detection rate, TPR and FPR. Figure 11 depicts a graphical representation of various models in which the proposed model outperforms at a greater rate. Figure 12 depicts a graphical representation of various models over computation time which will be obtained during the training period. Figure 13 depicts the output instances of segmentation.

Table 5: Overall analysis under detection rate, TPR, FPR.

Models	**Detection rate**	**TPR**	**FPR**
Alexnet	85	82	18
Googlenet	83	81	19
Inception v3	90	87	13
VGG19	86	83	17
Resnet50	78	73	27
VGG16	95	92	8

Figure 11: Models vs Measures. Overall analysis under detection rate, TPR and FPR

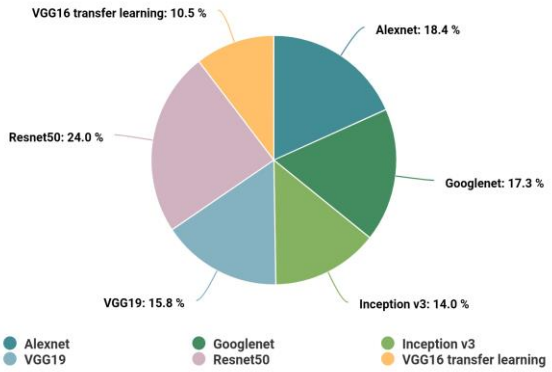
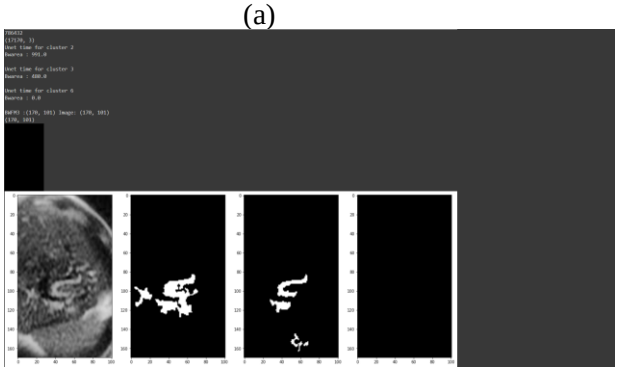
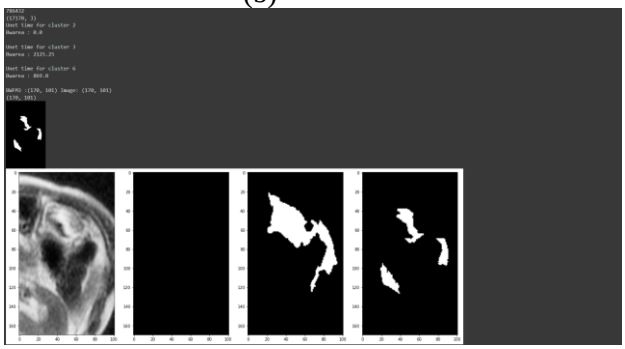


Figure 12: Models vs Computation time during the training period.

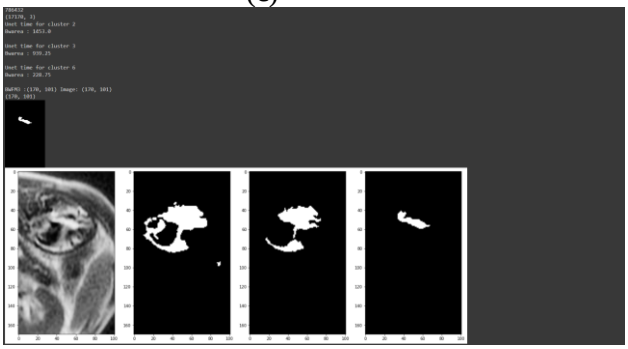
(a)



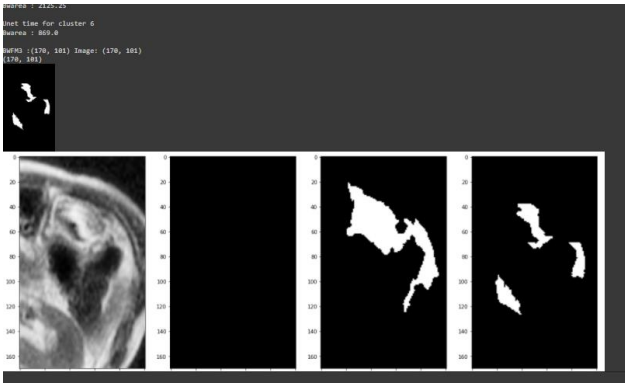
(b)



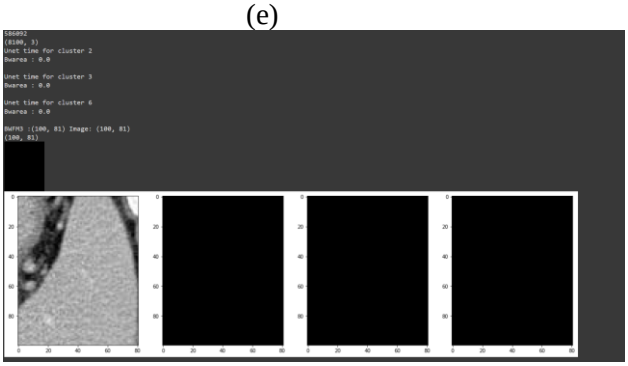
(c)



(d)



(e)



(f)

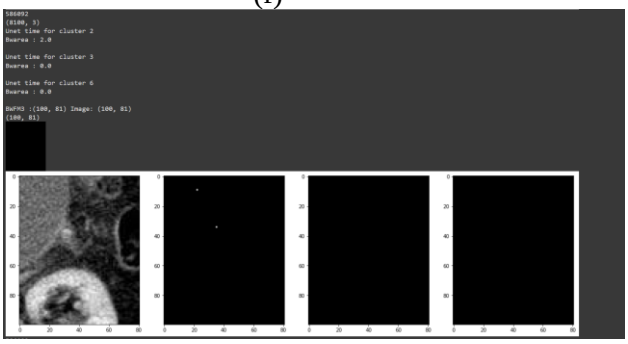


Figure 13. Segmentation output where (a,b,c) depict unhealthy output and (d,e,f) depicts the healthy output using UNET++.

5 Discussion

The purpose of this study is to demonstrate the effectiveness of MRI image analysis using the VGG-16 model with the Harris hawk's optimization (HHO) algorithm in segmentation and feature selection for pancreatic cancer classification from MRIs. Because MRI provides better contrast between fat, water, muscle, and other soft tissues than CT, it generally has a good spatial resolution compared to other modalities. Conventional MRI has shown a high degree of sensitivity and specificity for the detection of pancreatic tumors based on reviews of previous studies [41] with awareness of the presence of the tumor. The sensitivity of our proposed framework for the detection of pancreatic cancer was 96.34 on the test data set, as well as precision, recall, and F1 score that were considered as high compared to other approaches in the literature discussed in table 2. As a result, our

framework is comparable to the ability of humans to recognize images. In an analysis of 225 asymptomatic patients with a high risk of pancreatic cancer, Canto et al. (2017) [42] found that EUS (Endoscopic Ultrasound Scan) had the highest rates of tumor detection (42%) as compared to CT (11%), and MRI (33%). For tumor detection, Tonozoko et al. (2021) [17] used EUC imaging, which yields a higher sensitivity and can detect smaller tumors (Grade I). However, due to the risks and convenience issues associated with EUS, the MRI appears to be a better method. Hence in our model, we use MRI images for the detection of pancreatic cancer. The classification accuracy of the proposed method is 93.52 as compared to the image classification model of VGG-19 [18] which shows an accuracy of 87.52. According to Fu et al. (2021) [19], the inception model uses nuclear features that lead to false positive results that can be avoided by optimizing the selection of features. In our proposed model we utilized a VGG-16 framework with HHO-based CNNs and HHO-based Bags of visual terms for feature extraction and selection to improve the accuracy even with a smaller number of convolutional layers as compared to the VGG-19 model [18].

In general, unlike computers, the human brain does not perform at its best when fatigued, stressed, or limited in experience, which results in misdiagnosis or overlooking a lesion during an MRI. Artificial intelligence, on the other hand, can consistently provide reliable performance within a very short period, thereby compensating for the limitations of human capability and preventing human errors in clinical practice. As a result, our framework can be useful for both beginners learning MRI, as well as fatigued experts or carelessness caused by individuals who have accumulated a large number of screenings. Additionally, the data set for this study included a variety of images, including those with hazy borders and unclear images, which are frequently seen in clinical exams. These images were then enhanced using the Boosted Anisotropic Diffusion Filter (BADF) and Contrasted Limited Adaptive Histogram Equalization (CLAHE) algorithms to improve the image quality for better accuracy. Therefore, we believe that our system can detect diverse tumors by learning the images and through the utilization of better image-enhancing techniques and optimal feature selection strategies.

6 Conclusion

This paper brings an effective yet novel approach for pancreatic cancer detection at an earlier stage using deep learning. For this, initially, MRI data are collected from the popular repository CTAC-PDAC and with the help of CLAHE and BADF, preprocessing is done and then proceeded to segment cancer regions using UNet++. Further, for extracting quintessential features along with selection, the use of both HHO-based CNN and BOVW is done. Finally for effective use of transfer learning VGG16 is performed for detection. The proposed model outperforms better with 0.96% accuracy over state-of-the-art models under various measures. This paper will be helpful for another research specialist to dig deep and get to understand the stages and come up with better integrated and advanced models.

References

[1] National Cancer Institute (28 February 2021). Surveillance, Epidemiology, and End Results (SEER) Cancer Stat Facts: *Pancreatic Cancer*. Available online: https://seer.cancer.gov/statfacts/html/pancreas.html.

[2] Siegel, R.L.; Miller, K.D.; Jemal, A. (2020) Cancer statistics, *CA Cancer J. Clinicians*. 2020, 70, 7–30. https://doi.org/10.3322/caac.21590.

[3] Andersson.R., Vagianos.C.E.; Williamson.R.C. N (2004). Preoperative staging, and evaluation of resectability in pancreatic ductal adenocarcinoma. *HPB (Oxford)* 2004, 6, 5–12. https://doi.org/10.1080/13651820310017093

[4] Jonathan D Mizrahi, Rishi Surana, Juan W Valle, Rachna T Shroff (2020). Pancreatic cancer. *The Lancet*, 395, 2008–2020. https://doi.org/10.1016/s0140-6736(20)30974-0.

[5] Giada.M.M.(2020). The ambiguous role of the inflammatory micromilieu in solid tumors. *Pathology*.41,118–123. https://doi.org/10.1007/s00292-020-00837-1.

[6] Mayer, P.; Dinkic, C.; Jesenofsky, R.; Klauss, M.; Schirmacher, P.; Dapunt, U.; Hackert, T.; Uhle, F.; Hansch, G.M.; Gaida, M.M. (2018). Changes in the microarchitecture of the pancreatic cancer stroma are linked to neutrophil-dependent reprogramming of stellate cells and reflected by diffusion-weighted magnetic resonance imaging. *Theranostics*, 8, 13–30. https://doi.org/10.7150/thno.21089.

[7] Grosse-Steffen, T.; Giese, T.; Giese, N.; Longerich, T.; Schirmacher, P.; Hansch, G.M.; Gaida, M.M. (2012). Epithelial-to-mesenchymal transition in pancreatic ductal adenocarcinoma and pancreatic tumor cell lines: The role of neutrophils and neutrophil-derived elastase. *Clinical and Developmental Immunology*. 1-12. https://doi.org/10.1155/2012/720768

[8] Gaida, M.M.; Steffen, T.G.; Gunther, F.; Tschaharganeh, D.F.; Felix, K.; Bergmann, F.; Schirmacher, P.; Hansch, G.M. (2012). Polymer-photonuclear neutrophils promote dyshesion of tumor cells and elastase-mediated degradation of E-cadherin in pancreatic tumors. *Eur. J. Immunol*. 42, 3369–3380. https://doi.org/10.1002/eji.201242628.

[9] Verbeke, C.(2016). Morphological heterogeneity in ductal adenocarcinoma of the pancreas—Does it matter? *Pancreatology*. 16, 295–301. https://doi.org/10.1016/j.pan.2016.02.004.

[10] Hruban, R.H.; Adsay, N.V.; Albores-Saavedra, J.; Compton, C.; Garrett, E.S.; Goodman, S.N.; Kern, S.E.; Klimstra, D.S.; Kloppel, G.; Longnecker, D.S.; (2001). Pancreatic intraepithelial neoplasia: (A new nomenclature and classification system for pancreatic duct lesions). *Am. J. Surg. Pathol*.,25, 579–586. https://doi.org/10.1097/00000478-200105000-00003.

[11] Ren, B.; Liu, X.; Suriawinata, A.A. (2019) Pancreatic Ductal Adenocarcinoma and Its Precursor Lesions: Histopathology, Cytopathology, and Molecular Pathology. *Am. J. Pathol.*,189, 9–21. https://doi.org/10.1016/j.ajpath.2018.10.004.

[12] Esposito, I.; Hruban, R.H.; Verbeke, C.; Terris, B.; Zamboni, G.; Scarpa, A.; Morohoshi, T.; Suda, K.; Luchini, C.; Klimstra, D.S.; et al. (2020). Guidelines on the histopathology of chronic pancreatitis. Recommendations from the working group for the international consensus guidelines for chronic pancreatitis in collaboration with the International Association of Pancreatology, the American Pancreatic Association, the Japan Pancreas Society, and the European Pancreatic Club.*Pancreatology*,20, 586–593. https://doi.org/10.1016/j.pan.2020.04.009.

[13] Hanna, M.G.; Reuter, V.E.; Hameed, M.R.; Tan, L.K.; Chiang, S.; Sigel, C.; Hollmann, T.; Giri, D.; Samboy, J.; Model, C.; et al. (2019). Whole slide imaging equivalency and efficiency study: Experience at a large academic center. *Modern Pathology*.32, 916–928. https://doi.org/10.1038/s41379-019-0205-0

[14] Markl, B.; Fuzesi, L.; Huss, R.; Bauer, S.; Schaller, T. (2020). Number of pathologists in Germany: Comparison with European countries, USA, and Canada. *Virchows Arch*.478,2,335-341. https://doi.org/10.1007/s00428-020-02894-6

[15] Metter, D.M.; Colgan, T.J.; Leung, S.T.; Timmons, C.F.; Park, J.Y. (2019). Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. *JAMA Network Open*, 2, e194337. https://doi.org/10.1001/jamanetworkopen.2019.4337

[16] Jiang, Y.; Yang, M.; Wang, S.; Li, X.; Sun, Y. (2020). Emerging role of deep learning-based artificial intelligence in tumour pathology. *Cancer Commun. (Lond.)*, 40, 154–166. https://doi.org/10.1002/cac2.12012

[17] Tonozuka, R., Itoi, T., Nagata, N., Kojima, H., Sofuni, A., Tsuchiya, T., ... & Mukai, S. (2021). Deep learning analysis for the detection of pancreatic cancer on endosonographic images: A pilot study. *Journal of Hepato-Biliary-Pancreatic Sciences*, 28(1), 95-104. https://doi.org/10.1002/jhbp.825.

[18] Liu, K. L., Wu, T., Chen, P. T., Tsai, Y. M., Roth, H., Wu, M. S., ... & Wang, W. (2020). Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation. *The Lancet Digital Health*, 2(6), e303-e313. https://doi.org/10.1016/s2589-7500(20)30078-9

[19] Fu, Hao, et al. (2021). "Automatic pancreatic ductal adenocarcinoma detection in whole slide images using deep convolutional neural networks." *Frontiers in oncology*. 11. 2464. https://doi.org/10.3389/fonc.2021.665929

[20] Abbas, Sabah Khudhair, and Rusul Sabah Obied. (2021). "Novel Computer Aided Diagnostic System Using Synergic Deep Learning Technique for Early Detection of Pancreatic Cancer." *Webology 18. Special Issue on Information Retrieval and Web Search.* 367-379. https://doi.org/10.14704/web/v18si02/web18105

[21] Li, J., Qi, L., Chen, Q., Zhang, Y. D., & Qian, X. (2022). A Dual Meta-Learning Framework based on Idle Data for Enhancing Segmentation of Pancreatic Cancer. *Medical Image Analysis*,78. 102342. https://doi.org/10.1016/j.media.2021.102342

[22] Online source: https://www.google.com/url?sa=t&source=web&rct=j&url=https://wiki.cancerimagingarchive.net/plugins/servlet/mobile%3FcontentId%3D21267608%23content/view/21267608&ved=2ahUKEwjR0tC2-KH3AhVUS2wGHcYYCBAQFnoECA4QAQ&usg=AOvVaw0pVpGZHP1Z40YdRfI5vBnt.

[23] Suman, G., Patra, A., Korfiatis, P., Majumder, S., Chari, S. T., Truty, M. J., ... & Goenka, A. H. (2021). Quality gaps in public pancreas imaging datasets: implications & challenges for AI applications. *Pancreatology*, 21(5), 1001-1008. https://doi.org/10.1016/j.pan.2021.03.016

[24] Rao, K., Bansal, M., & Kaur, G. (2022). Retinex-Centred Contrast Enhancement Method for Histopathology Images with Weighted CLAHE. *Arabian Journal for Science and Engineering*, 47(11),13781-13798. https://doi.org/10.1007/s13369-021-06421-w

[25] Rodríguez-Pena, A., Uranga-Solchaga, J., Ortiz-de-Solórzano, C., & Cortés-Domínguez, I. (2020). Spheroscope: A custom-made miniaturized microscope for tracking tumour spheroids in microfluidic devices. *Scientific Reports*, 10(1), 1-12. https://doi.org/10.1038/s41598-020-59673-1

[26] Sawssen, B., Okba, T., & Noureeddine, L. (2022). A mammographic image classification technique via the Gaussian Radial Basis Kernel ELM and KPCA. Conference:*International conference on Mathematics and Computers in Science and Engineering*, at H10,spain.

[27] Uplaonkar, D. S., & Patil, N. (2021). Ultrasound liver tumor segmentation using adaptively regularized kernel-based fuzzy C means with the enhanced level set algorithm. *International Journal of Intelligent Computing and Cybernetics*. 15(3),438-453.https://doi.org/10.1108/ijicc-10-2021-0223

[28] Iima, M., & Le Bihan, D. (2016). Clinical intravoxel incoherent motion and diffusion MR imaging: past, present, and future. *Radiology*,

278(1), 13-32. https://doi.org/10.1148/radiol.2015150244

[29] Goyal, B., Dogra, A., Agrawal, S., Sohi, B. S., & Sharma, A. (2020). Image denoising review: From classical to state-of-the-art approaches. *Information fusion*, 55, 220-244. https://doi.org/10.1016/j.inffus.2019.09.003

[30] Long, J., Shelhamer, E., Darrell, T. (2015): Fully convolutional networks for semantic segmentation. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440. https://doi.org/10.1109/cvpr.2015.7298965

[31] Ronneberger, O., Fischer, P., Brox, T. (2015): U-Net: convolutional networks for biomedical image segmentation. *LNCS*, vol. 9351, pp. 234–241. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[32] Zhang, L., Shi, Y., Yao, J., Bian, Y., Cao, K., Jin, D., ... & Lu, L. (2020). Robust pancreatic ductal adenocarcinoma segmentation with multi-institutional multi-phase partially-annotated CT scans. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 491-500 https://doi.org/10.1007/978-3-030-59719-1_48

[33] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3-11). https://doi.org/10.1007/978-3-030-00889-5_1

[34] Basha, J., Bacanin, N., Vukobrat, N., Zivkovic, M., Venkatachalam, K., Hubálovský, S., & Trojovský, P.(2021). Chaotic Harris hawks optimization with quasi-reflection-based learning: An application to enhance CNN design. *Sensors*, 21(19), 6654. https://doi.org/10.3390/s21196654

[35] Thaher, T., Heidari, A. A., Mafarja, M., Dong, J. S., & Mirjalili, S. (2020). Binary Harris Hawks optimizer for high-dimensional, low sample size feature selection. *Algorithms for Intelligent Systems.*251-272. https://doi.org/10.1007/978-981-32-9990-0_12

[36] Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., & Greenspan, H. (2018). Chest pathology identification using deep feature selection with non-medical training. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3), 259-263. https://doi.org/10.1080/21681163.2016.1138324

[37] Chaib, S., Gu, Y., & Yao, H. (2015). An informative feature selection method based on sparse PCA for VHR scene classification. *IEEE Geoscience and Remote Sensing Letters*, 13(2), 147-151. https://doi.org/10.1109/lgrs.2015.2501383

[38] Huang, M., Yang, W., Yu, M., Lu, Z., Feng, Q., & Chen, W. (2012). Retrieval of brain tumors with region-specific bag-of-visual-words representations in contrast-enhanced MRI images. *Computational and mathematical methods in medicine*.1-17. https://doi.org/10.1155/2012/280538

[39] Baldota, S., Sharma, S., & Malathy, C. (2021, July). Deep Transfer Learning for Pancreatic Cancer Detection. *In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* 1-7. https://doi.org/10.1109/icccnt51525.2021.9580000

[40] Sehmi, M. N. M., Fauzi, M. F. A., Ahmad, W. S. H. M. W., & Chan, E. W. L. (2021). Pancreatic cancer grading in pathological images using deep learning convolutional neural networks. *F1000Research*, 10(1057), 1057. https://doi.org/10.12688/f1000research.73161.1

[41] Costache, M. I., et al (2017). "Which is the best imaging method in pancreatic adenocarcinoma diagnosis and staging-CT, MRI or EUS?" *Current health sciences journal* 43(2).132-136. http://dx.doi.org/10.12865/CHSJ.43.02.05

[42] Canto, Marcia Irene, et al. (2012). "Frequent detection of pancreatic lesions in asymptomatic high-risk individuals." *Gastroenterology* 142(4) 796-804. https://doi.org/10.1053/j.gastro.2012.02.029

Assessing Mental Health Crisis in Pandemic Situation with Computational Intelligence

Megha Rathi, Adwitiya Sinha*, Siddhant Tulsyan, Avishka Agarwal, Anushka Srivastava
Dept. of Comp. Sc. & Engineering and Information Technology Jaypee Institute of Information Technology, India
E-mail: megha.rathi@jiit.ac.in, mailtoadwitiya@gmail.com, siddhant05tulsyan@gmail.com, avishka2404@gmail.com, anushka.srivastava2398@gmail.com
*Corresponding author

Keywords: mental health crisis, healthcare information management, computational intelligence, machine learning, synthetic minority oversampling, covid-19, biomedical informatics.

Received: January 6, 2022

The coronavirus pandemic has created huge emotional distress and increased the risk of psychiatric problems. This happened owing to imposition of necessary stringent healthcare measures that infringed personal space, emotional freedom, and caused financial loss. Our physical well-being is directly associated with mental fitness and health. From analysis it has been found that feature like struggling in concentration and memory, visionary issues, and arthritis are customary symptoms in patients suffering from mental crises. Our proposed research work aims to find out the reasons behind mental illness and ways to improve mental disorders using supervised approach. The main focus is to develop a smart computationally intelligent model to assist healthcare practitioners in analysing and diagnosing severe mental illness. Our proposed model assists in analysing causes of mental disorder and aids in reducing total medicinal cost along with reduced mental illness rate. Additionally, a recommendation system is also developed for diagnosing depressive patients.

Povzetek: Opisana je inteligentna metoda za pomoč pri mentalnih boleznih, povezanih s pandemijami.

1 Introduction

The public health emergencies imposed during Covid-19 pandemic has caused distressed in communities at large. The mandating of sudden and unfamiliar public safety norms have caused emotional distress among people [1]. As the normal course of living was severely encroached by home confinement and social distancing, many cases of mental health crisis started to erupt. Moreover, people who suffered from recurrent ailments during the pandemic become even more vulnerable to psychiatric problems and other severe health havocs. As a result, the yearly medicinal cases for mental disorders started increasing globally, and hence it become essential to reveal the root causes for mental disorders, including anxiety, depression, and many more adverse psychosocial disorders [2]. Moreover, the total expenditure for treating patients also increased, which includes restorative cost for treatment. In order to understand these ballooning costs, several large-scale epidemiological studies are being conducted to provide information on the health of United States citizens. One such study, the Behavioural Risk Factor Surveillance System (BRFSS), conducts surveys to collect uniform data on health risk behaviours, chronic diseases, access to healthcare information, and to employ preventative medical services in the United States [3]. This survey provides valuable information on behavioural patterns which, if coupled with current big data and machine learning techniques, may help to provide valuable insights into persons at risk of mental health crises. By targeting and understanding these populations, preventative health measures could be put into place to ultimately help lower health care costs in the United

States. Adults with depression and anxiety are significantly more expected to smoke, to be obese, to be physically inactive, to binge drink, and to drink more heavily than those who do not display any symptoms of depression and anxiety. Additionally, a dose-dependent relation exists between severity of depression and the smoking intensity, obesity, and physical inactivity, in which individuals who are more depressed become prone to heavy engagements in such activities. In a study of the 2012 Behavioural Risk Factor Surveillance System (BRFSS) data, found that there are significant relationships between depression and childhood mental illness, limited usual activity, and abuse [4]. In proposed research work J48 classification tree is used to predict depression with 82% accuracy, using these predictive attributes. Our research aims to create a solid foundation with the use of machine learning in helping to predict mental crises using multiple health attributes.

2 Background study

Extensive research and case studies were conducted in assessing the acuteness of emotional distress and forecasting mental health crisis. The authors in [5] have thoroughly discussed on the major stressors caused due to

quarantine and isolation measures, and different ways to reduce its impact. In [6], several tools and measures were suggested for measuring the psychological impact of the Covid-19 pandemic. Moreover, technicians often face scarcity and imbalance in healthcare data that pose a major challenge for training models and supervised learning. This has been taken forward by the author in [7] to deal with the development of classifiers from imbalanced datasets. A dataset is considered to be imbalanced, when the characterization classes are not roughly similar. Frequently certifiable informational indexes are predominately made out of ordinary precedents with just a little level of strange or intriguing models. It is additionally the situation that the expense of misclassifying an anomalous (fascinating) model as an ordinary precedent is regularly a lot higher than the expense of the invert blunder. The authors have demonstrated that their proposed technique can accomplish better classifier execution for over-examining the minority (strange) class and under inspecting the greater part (typical) class in the Receiver Operating Characteristic (ROC) space, than just under testing the larger part class. In another novel work, Synthetic Minority Oversampling Technique (SMOTE) Rough Set Theory (RST)is proposed, which is dependent on oversampling and under sampling for high imbalanced informational indexes [8]. SMOTE-RSB is a hybrid data pre-processing approach that manages imbalanced informational indexes through the development of new examples and samples, utilizing SMOT together with the use of an altering method dependent on the RST and the lower estimation of a subset. The proposed technique has been approved by a trial think about demonstrating great outcomes utilizing C4.5 as the learning calculation.

Multi-mark learning has been turning into an inexorably dynamic region into the machine learning group since a wide variety of true issues are normally multi-named. Destroyed is an oversampling system that has been effectively connected for adjusting single-marked informational indexes, however has not been utilized in multi-name structures up until now. In this regard, authors in [9] highlighted a few methodologies are proposed and contrasted by the author all together with produce manufactured examples for adjusting informational indexes in the preparation of multi-name calculations. Results demonstrate that a right determination of seed tests for oversampling improves the grouping execution of multi-mark calculations. In yet another novel work [10], authors inspected the general social insurance costs related with sorrow depression and also, tension among essential consideration patients. Out of 2110 back to back essential consideration patients in a wellbeing support association, 12-thing Health General Questionnaire were screened with 1,962 people. 615 patients were further selected for indicative appraisal; Composite International diagnostic review performed on 373 patients and 328 were re-examining 12 months after the fact. Electronic cost records were utilized to compute absolute human services costs for the half year time frame encompassing the gauge evaluation and a comparative period encompassing the subsequent appraisal. Cost contrasts reflected higher use of general therapeutic administrations instead of higher psychological wellness treatment costs. In research [11], authors used computerized record frameworks of a vast staff model well-being up keep association (HMO) were utilized to distinguish sequential essential consideration patients with visit findings of sorrow and a correlation test of essential consideration patients with no melancholy conclusion. Comparable cost contrasts were watched for every one of the subdivisions inspected (treatment using antidepressants, treatment without antidepressants, and patients analysed at routine physical clinical visits). Drug store records showed more noteworthy perpetual medicinal sickness in the analysed discouragement gathering, however huge cost contrasts stayed after alteration ($3971 versus $2644). Two overlap cost contrasts endured for no less than a year after commencement of treatment. As an end, creators found that finding of misery is related with a summed-up increment being used of wellbeing administrations that is just halfway clarified by co grim ailments.

The authors of the paper [12], regulated a poll to 367 patients with type-1 and type-2 diabetes from the primary care clinics of two healthcare information management organizations, to get information on socioeconomics, burdensome side effects, diabetes learning, working, and diabetes self-care. Based on computerized information, we quantified therapeutic comorbidity, social insurance costs, glycosylated haemoglobin (HbA1c) levels, and oral hypoglycaemic remedy refills. Utilizing burdensome side effect seriousness tertiles (less, mid-range, or highest), they performed relapse investigations to decide the effect of burdensome indications on constancy to diabetes self-support and oral hypoglycaemic regimens, HbA1c levels, utilitarian debilitation, and human services costs. Compared with patients in the low-seriousness gloom side effect tertile, those in the medium and high-seriousness tertiles were essentially less follower to dietary suggestions. Further investigations testing the viability and cost-adequacy of upgraded models of consideration of diabetic patients with sorrow are required. In yet another contribution in the field of mental illness authors have provided information about imbalanced learning issues that hold an unlike conveyance of information tests among various classes and represent a test to any classifier as it turns out to be difficult to get familiar with the minority class tests [13]. This paper distinguishes that the majority of the current oversampling techniques may create the wrong engineered minority tests in certain situations and make learning undertakings harder. To overcome this, Majority Weighted Minority Oversampling Technique (MWMOTE) is introduced for productively handling with variant learning issues. MWMOTE first distinguishes the difficult to-learn educational minority class tests and relegates them loads as per their Euclidean separation from the closest larger part class tests. In another contribution, the authors have shown a novel Cluster Based Synthetic Oversampling (CBSO) algorithm in the proposed study [14]. CBSO receives its fundamental thought from existing manufactured oversampling methods and consolidates unsupervised clustering in its

engineered information age system. One of the core machine learning algorithms that gained achievement in health analytics is Support Vector Machine (SVM). Statistics of SVM makes it suitable to handle all type of medical datasets. In numerous settings, we additionally have the choice of utilizing pool-based dynamic learning. Dynamic Learning with help vectors is examined in the study [15], i.e., a computation for picking which examples to demand straightaway. In another work, comparative

Table 1: Summarized application of machine learning techniques in mental health analysis.

S.No.	Author, Year	Objective	Approach	Results
1.	O. Oyebode, F. Alqahtani and R. Orji, 2020 [24].	In the recent study authors have analyzed mental health apps. They have evaluated online available 104 mental health apps and perform sentiment analysis on reviews.	Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Random Forest (RF).	F1 Score and accuracy is compared and it is found that SGD achieved the best overall F1 score of 89.42 then followed by SVM, and LR.
2.	Ela Gore, Sheetal Rathi, 2019 [25].	In this work, authors surveyed researches done for the applicability of machine learning for mental heal analysis.	This paper surveyed numerous machine and deep learning models as SVM, K-Nearest Neighbor (KNN), Random Tree, Convolution Neural Network (CNN), Recurrent Neural Network (RNN) etc.	From the survey it is concluded that SVM with their different kernels and CNN models utilized in many of the research work. They also give better results in terms of parameters like accuracy, etc.
3.	Sabourin, A. A., Prater, J. C., & Mason, N. A., 2019 [26].	In today's competitive era students are at high mental health risk. Authors compared the mental health status of pharmacy students to other university students.	Computational techniques like SVM, Naive Bayes (NB), KNN, and Random Forest (RF) used.	RF achieves precision approximately equal to 83.33%, NB 71.42%, SVM 85.71% and KNN 55.55%.
4.	Hou, Y., Xu, J., Huang, Y., & Ma, X. ,2016 [27].	This one is another significant work done for analyzing mental health profile of students. It targets to find association between reading habits of students and depression induced due to reading	Compare algorithms like SVM, KNN, Decision Tree DT, Artificial Neural Network (ANN) , and Bayesian Classifier.	Most Accurate classifier is SVM with 82% accuracy.
5.	Gokten, E. S., & Uyulan, C. ,2021 [28].	Advanced machine learning techniques are applied to predict psychiatric disorders	Random Forest is used and applied on a record of 482 children.	Following results were obtained for kids with mental disorder: Accuracy= 72%, F1-Score=71%, Precision= 72%, and Recall= 71%.
6.	Xin, Y., Ren, X, 2022 [29].	Purpose of this research work is to forecast the psychiatric illness amongst old age people from the aspects like health profile, relationship with family, social behaviour, demographic location, and behaviour of health.	This paper used the random forest classifier to predict the depression of old age people.	The psychiatric disorder of rural old age grouped was 57.67%, and that of urban was 44.59%.
7.	Srividya, M., Mohanavalli, S., & Bhalaji, N., 2018 [30].	Application of numerous machine learning techniques to identify mental health is the main objective of this work.	Logistic Regression (LR), SVM, NB, DT, KNN, RF, and Bagging.	Highest Accuracy achieved by ensemble approach Bagging (90%) and RF (90%) followed by SVM (89%) and KNN (89%).
8.	Tate, A. E., McCabe, R. C., Larsson, H., Lundström, S., Lichtenstein, P., & Kuja-Halkola, R., 2020 [31].	A Machine Learning Model is developed and compared for predicting mental illness in adolescence. All techniques are explored based on statistical evaluation parameters.	RF, XGBoost, Neural Network (NN), logistic regression (LR), neural network and SVM.	Models compared using Area under Curve (AUC) and it is noticed that SVM and RF had highest AUC's equals to 0.754.

9.	Reddy, U. S., Thota, A. V., & Dharun, A., 2018 [32].	Stress patterns are analyzed in working professionals using machine learning techniques in order to highlight the factors that strongly affect the stress level.	LR, KNN, DT, Boosting, Bagging, RF.	From the results it has been concluded that embedded approach boosting achieves highest 75.13% accuracy.

analysis of various computational intelligence mathematical statistics for various infections determination, for instance, heart disease, diabetes, dengue, and hepatitis is presented [16]. Main emphasis of this review work to highlight the importance of machine learning techniques towards decision support system and diagnostics. In yet another novel work, authors highlighted the major mental issues also explored treatment coverage country wise [17]. In another survey [18] author has cited the importance and significance of smart devices for assessing anxiety, stress, and depression. Various work has been done in the area of health informatics for finding and extracting valuable insights using machine learning techniques [19-20]. From these researches it has been concluded that machine learning plays significant role in extracting and predicting health outcomes [21-23, 41-43].

In our research initiative, supervised machine learning approach is used to build a computationally efficient model to serve the mental health crisis in the society. Our proposed model ensures biomedical applicability by aiding the doctors to provide reliable healthcare service delivery to patients with mental health issues. List of related work in the domain of analyzing mental health illness is presented in Table I.

3 Proposed methodological framework

In our research, the BRFSS dataset was considered, which further required downstream analysis. This required data scrubbing and pre-processing techniques for cleaning and preparing the data for experimentation. Various machine learning algorithms were applied on the cleaned data set and respective accuracies were predicted. Recommendation system was built on the basis of this model using shiny web application [33-34].

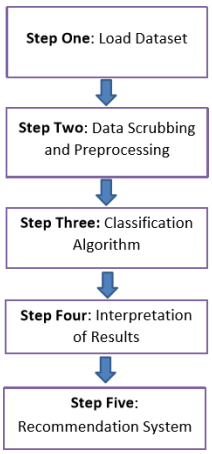


Figure 1: Structural Flow of Proposed Framework.

3.1. Data collection

The Behavioural Risk Factor Surveillance System (BRFSS) is a random annual phone-based survey which tracks health risk behaviours, chronic diseases, access to health care, and the use of preventative healthcare service management in the United States, available freely for access [4]. The most current data year (2016) was used for this project, which contained 450 attributes and 486,303 records. All questions asked in the survey (attributes) are available in [3]. Mental illness was characterized by individuals who had current anxiety and depression, life-time depression detection, and or a lifetime anxiety diagnosis and the class attribute (Mental Crises) were compiled based on these answers.

3.2. Data processing & scrubbing

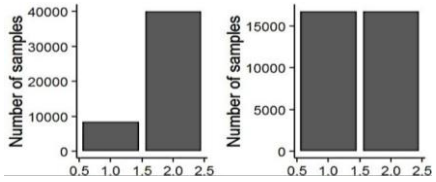
Data scrubbing is the necessary action required to remove repeated, incorrect, and improperly data from the dataset [35-37]. We renamed data frame to prevent overwriting the original file, and identify the column names.

The attributes of original data were written in their short forms which were not easy to comprehend. These attribute names were expanded to make more sense of the data. It helped to read the data easily and connect different habits of a patient with its mental status. Since there were 450 attributes, some of these attributes were removed which were not needed Attributes that had no relevant meaning or no practical significance like telephone number, address, number of family members, etc that summed up to 60 columns, were removed. Record identification column was removed from the data base as it is unnecessary for downstream analysis. Our dataset consists of 6, 17, 07, 536, and NA values. This value was quite huge and hence was interfering in the various machine learning algorithms. Survey contained answer choices in the form of *none* (88), *do-not-know* (7), *refused* (9), etc. which were replaced to NA as it did not contribute in prediction. To normalize the data set, all the NA values were then replaced by means of their respective columns. Several attributes were explored. Count of *no* and *yes* was checked in the output column (*depressive*). This was done to check the proportion of *no* to *yes*. The ratio came out to be 1:4. Due to the less count of *no*, model prediction was not very accurate. Since data was quite huge so due to computational limitations, data set was sub sampled to 10% of the original data set. We made sure ratio of *no*to *yes* does not change in the sub sampled data, suggesting the smaller data set is representative of the whole data set. Data Scrubbing also included removing incomplete attributes

(i.e. those with >25% unanswered answers) and transforming attributes for downstream processing.
Data pre-processing is applied to transform raw data into a format that is easily understandable and upgrade the classifier performance [38]. Synthetic Minority Over-sampling Technique (SMOTE) was used to combat an imbalanced class design and to maintain the yes to no ratio in the sub sampled dataset. Fig 2 shows the comparison of number of classes (*yes* and *no*) before and after SMOTE. This strategy enables us to adjust the class configuration, wiping out any predisposition that may ruin our downstream analyses. Unbalanced classification issues cause problems to many learning calculations and algorithms. These issues are portrayed by the uneven extent of cases that are accessible for each class of the issue. SMOTE is a notable calculation to tackle this issue. Moreover, the dominant part class precedents are additionally under-examined, prompting an increasingly adjusted dataset.

The parameters *perc.over* and *perc.under* control the measure of over examining of the minority class and under-sampling of the majority classes, respectively. *perc.over* will typically be a number over 100. With this kind of qualities, for each case in the dataset having a place with the minority class, new instances of that class were made. In the event that *perc.over* is an incentive underneath 100 than a solitary case will be created for a haphazardly chosen extent (given by *perc.over*/100) of the cases having a place with the minority class on the first informational collection. The parameter *perc.under* controls the extent of instances of the dominant part class that will be arbitrarily chosen for the last *adjusted* informational index. This extent is determined as for the quantity of recently created minority class cases. The parameter k controls the manner in which the new precedents or examples are made. These precedents will be produced by utilizing the data from the k-nearest neighbours of every case of the minority class. The parameter k controls what number of these neighbours are utilized. This produces an arbitrary arrangement of minority class perceptions, utilizing bootstrapping and the datum point having k-closest neighbours. This decreased the predisposition towards the larger part class, while guaranteeing the new examples in the minority class were illustrative of the previous qualities. In this capacity k was set to be 5 and *perc.over* to be 110. The figure (2) demonstrates beginning number of *no*, which were 10000; while that of *yes* were 40000. Subsequent to applying SMOTE, number of *no* expanded to 18000 and *yes*, diminished to 12000.To further clean the dataset, Pearson correlation test was used to determine the correlation between each feature and the class attribute. Attributes with less than 10% correlation were discarded from downstream analysis.

Pearson correlation test. This is an estimate of precise association between two given variables of a system. Pearson correlation coefficient (r) is an estimate of the strength of the connection between the two variables. It has a value ranging from [-1,1]. If both variables increase and decrease together it implies positive correlation while if the value of one variable decrease with the increase in other variable value or vice-versa it indicates negative correlation.



(a) (b)

Figure 2: Comparison of number of classes (*yes*-*no*) (a) before and (b) after applying SMOTE.

3.3. Data classification

After the dataset was pre-processed and cleaned, machine learning algorithms were applied to examine its accuracy. Supervised algorithms such as k-nearest Neighbour, Random Forest, Decision Tree and SVM were applied. SVM gave an accuracy of 65% while KNN gave a precision of 70.16%. Random forest achieved an average accuracy score of 80% when *n_tree* was set to 100. Best accuracy was achieved through Decision Tree which gave 81.19% precision. A description of confusion matrix is given in Table 3. The decision tree was assembled utilizing 10-fold cross validation. The picked calculation, C4.5 or J48, was built utilizing a multistep process is presented in Table II. To start with, the single variable was discovered which best parts the information into two groups. Second, the information was separated, and the procedure was rehashed recursively until the subgroups either achieved a greatest size of 5 or no further modifications were made.

This methodology utilized a splitting criterion known as the *gain-ratio*, and was pruned utilizing a bottom up system known as *error-based* pruning. At last, precision and Area under the Curve (AUC) was surveyed to decide the reliability of the last tree and model. The Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) is a decent measure of the execution of a model. The AUC esteem can go from 0.5 (the model plays out no superior to arbitrary shot) to 1 (model suitably clarifies the reaction inside the test set).

3.4. Building recommendation model

A recommendation system was compiled to provide a user-interface program for use by doctors when their patients are in the examination room. We have developed this interface using shiny web application. This visualization helped us to give some insights on how habits like smoking, sleeping, remembering, etc can affect their mental health. All the responses of the user are recorded and scaled. We selected 6 questions according to highest gain ratio that were achieved in our decision tree model. These questions are illustrated as following.

- Have you visited a doctor for routine check-ups in last 6 months?
- Do you have memory loss issues, Concentration Issues, or Trouble in finalizing decisions?
- Do you have diabetes?
- Medical history of disease like: arthritis, lupus, fibromyalgia, or gout.
- Do you have any visionary impairment?
- Details of health policies of patient. Whether person is under health cover or not?

These questions were clustered further with other six questions whose correlation coefficient came out to be more that 10%. Response to every question was grouped with these questions to give an average depressive score. If the average depressive score is more than 50% then it represents that population in this cluster is more likely to be depressive.

Table 2: Class-view & multiple attributes in mental health dataset.

S.No.	Attribute	Values	Correlation with class attribute
1.	General Health	1. Excellent; 2. Very Good; 3.Good; 4. Fair; 5. Poor	-0.295607016
2.	Multiple Healthcare Professionals	1. Only one; 2.More than one; 3. None	0.103815018
3.	Cost prohibiting seeing a doctor	1. Yes; 2. No	0.165653515
4.	Participate in physical activities or exercise in past month	1. Yes; 2. No	-0.158701513
5.	Having disease Asthma	1. Yes; 2. No	0.100175029
6.	Having disease COPD	1. Yes; 2. No	0.186787737
7.	Having disease Arthritis	1. Yes; 2. No	0.237111272
8.	Time of last visit to dentist/ dental clinic	1. within the year; 2.within past 2 years; 3. within past 5 years; 4. five or more years ago	-0.149358224
9.	Number of permanent teeth removed	1. 1-5; 2. 6 or more; All; 4. None	0.139458736
10.	Gender of Respondent	1. Male; 2. Female	-0.248546199
11.	Marital status	1. Married; 2. Divorced; 3. Widowed; 4. Separated; 5. Never Married	-0.136709092
12.	Education level	1. Never attended; 2. Elementary; 3. Some High School; 4. High School Graduate; 5. Some College/ Technical; 6. College Graduate	0.118850168
13.	Own/Rented home	1. Own; 2. Rented; 3. Other Arrangement	-0.251743782
14.	Employment status	1. Employed; 2. Self-Employed; 3. Out of work for more than one year; 4. Out of work for less than one year; 5. Home maker; 6. Student	-0.38432588
15.	Blind/difficulty in seeing	1. Yes; 2. No	0.251269243
16.	Difficulty in remembering/concentrating	1. Yes; 2. No	0.442148749
17.	Difficulty walking/climbing stairs	1. Yes; 2. No	0.215583219
18.	Difficulty dressing/bathing	1. Yes; 2. No	0.124759584
19.	Difficulty doing errands alone	1. Yes; 2. No	0.245702066
20.	Smoked at least 100 cigarettes in entire life	1. Yes; 2. No	0.109147887
21.	Frequency of days currently smoking in month	1. Every day; 2. Some days; 3. Not all days	0.133169638

22.	Have delayed getting medical care	1. Yes; 2. No	0.16235243
23.	Been without healthcare services in past 12 months	1. Yes; 2. No	0.184359727
24.	Activity has been limited due to health problems	1. Yes; 2. No	0.191887956
25.	Having health problems that require special equipment	1. Yes; 2. No	0.1665181842
26.	Been diagnosed with depressive disorder (class attribute)	1. Yes; 2. No	1

4 Results & discussions

The correlation values between various attributes and 'depressive' show common symptoms that a patient might be dealing with in mental crises during the pandemic lockdown [39]. The results for symptoms, including difficulty in concentrating or remembering, blindness and arthritis is shown in figures 5-7. These symptoms are quite common in a person suffering from a mental crisis. Correlation coefficients of these attributes were 0.442148749, 0.251269243 and 0.215583219 respectively. Further, in figure 8, the relationship between depressive and health coverage is illustrated.

Also, we have compared our results from other existing work in the same domain listed in Table II. It has been found that the discoveries of this model help the consequences of past examinations, emphatically connecting burdensome scatters and dimensions of periodontal ailment, and proposing a negative connection with tooth brushing and dental checkups to melancholy may exist. No other existing work finding out the correlation amongst attributes as we did in the proposed work which is quite effective in highlighting the positive and negative features that directly or adverse impact the outcome. Best accuracy was achieved through Decision Tree, which gave 81.19% precision. The true positives ratio came out to be 34.9 while true negatives ratio was 46.2. This low FN rate is basic in a working model, as the cost of misclassifying a mental disease is a lot higher than the expense of misclassifying a non-mental disease. The highlights with the most elevated data gain give intriguing bits of knowledge into the respondents' practices in this investigation.

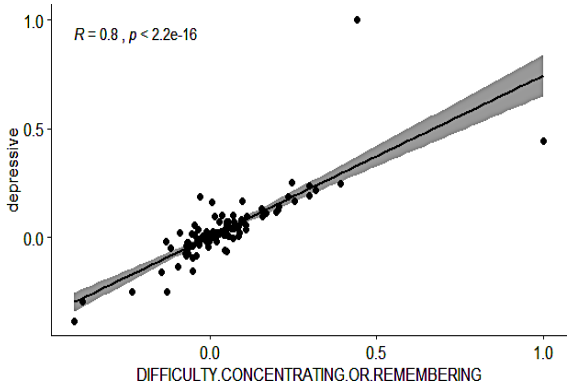


Figure 5: Relationship between depression& difficulty in concentration.

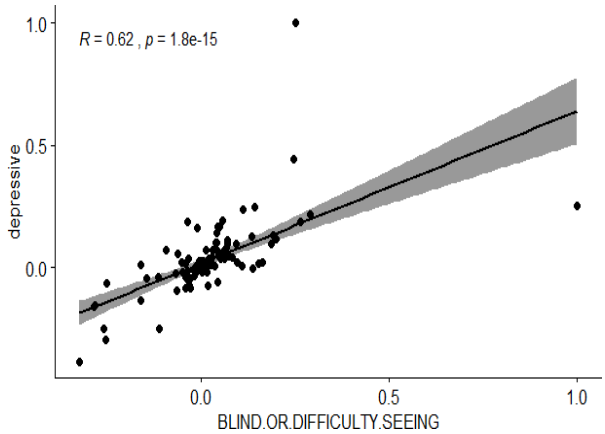


Figure 6: Relationship between depression& blindness.

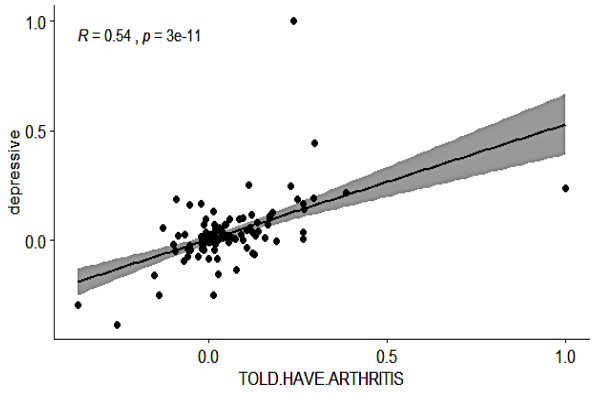


Figure 7: Relationship between depression& arthritis.

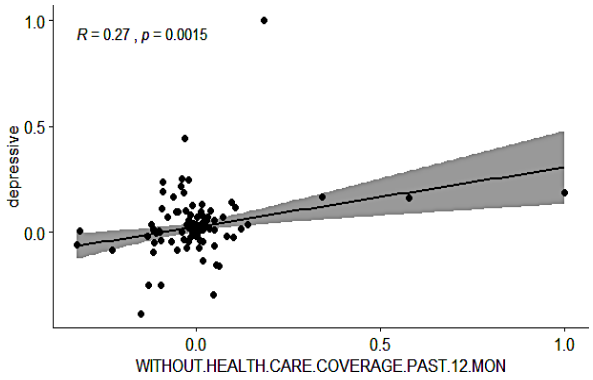


Figure 8: Relationship between depression & health coverage.

Table 3: Confusion matrix for proposed mental health model

	yes	***no***
yes	34.9	3.8
no	15.1	46.2

For our proposed model, the AUC was 0.83 as shown in figure 9. As it can be seen, the flat line initially depicts bad precision of the model. As soon as specificity value reaches a certain value of around 0.6, it escalates to a maximum value of 0.83. This shows the model has reached its maximum accuracy and hence it becomes constant thereafter. The decision tree of observed parameters is highlighted in figure 10.

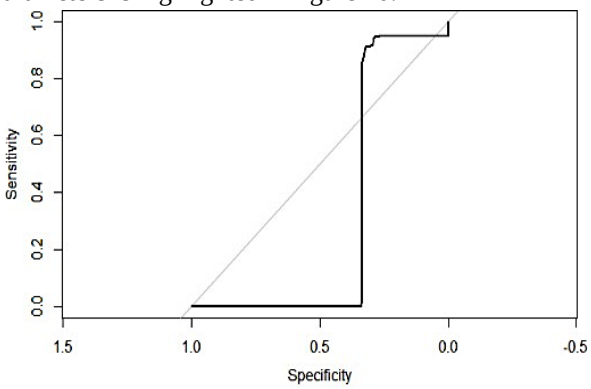


Figure 9: AUC of receiver operating characteristic.

```
DIFFICULTY.CONCENTRATING.OR.REMEMBERING = 1
|  WITHOUT.HEALTH.CARE.COVERAGE.PAST.12.MON < 2
|  |  BLIND.OR.DIFFICULTY.SEEING < 1.95 : 1 (2472/122)
|  |  BLIND.OR.DIFFICULTY.SEEING >= 1.95 : 1 (2687/786)
|  WITHOUT.HEALTH.CARE.COVERAGE.PAST.12.MON >= 2
|  |  TOLD.HAVE.ARTHRITIS < 2 : 1 (4319/452)
DIFFICULTY.CONCENTRATING.OR.REMEMBERING = 2
|  LAST.VISITED.DENTIST.OR.DENTAL.CLINIC < 1
|  |  LENGTH.OF.TIME.SINCE.LAST.ROUTINE.CHECKU < 1 : 2 (12737/4202)
|  |  LENGTH.OF.TIME.SINCE.LAST.ROUTINE.CHECKU >= 1 : 2 (4219/1904)
|  LAST.VISITED.DENTIST.OR.DENTAL.CLINIC >= 1
|  |  X.EVER.TOLD..YOU.HAVE.DIABETES = 2 : 1 (3033/742)
|  |  X.EVER.TOLD..YOU.HAVE.DIABETES = 1 : 2 (7907/3903)
```

Figure 10: Truncated decision tree outcome.

A confusion matrix summarizes the performance of the model [40]. The confusion matrix for Decision Tree is presented in Table III, and the model accuracy (calculated as (true observations/all observations)) was 81.07%. Table IV presented all the precision scores in descending order. From the table it is clear that Decision Tree outperform all other algorithm.

Table 4: Comparative analysis of precision score

Algorithm	**Precision Score**
Decision Tree	81.1935%
Random Forest	80.1265%
KNN	70.6312%
SVM	65.3542%

5 Conclusion & future research directions

Our research initiative addresses the ever-increasing crisis surfacing due to mental health related ailments, especially in Covid-19 pandemic situation. A set of supervised algorithms, including K-nearest Neighbor, Random Forest, Decision Tree and SVM were applied. Our proposed framework based off this model can help biomedical specialists in rapidly distinguishing in danger patients, prompting both higher rates of precaution medicinal services and early intercession, at last bringing down social insurance costs related with treating discouragement and tension in the country. Future undertakings should concentrate on expanding generally speaking exactness of the model to guarantee unwavering quality while giving specialists course with respect to their emotional well-being patients.

6 References

[1] Pfefferbaum, Betty, and Carol S. North. "Mental health and the Covid-19 pandemic." New England Journal of Medicine 383, no. 6 (2020): 510-512.

[2] Schäfer, Sarah K., M. Roxanne Sopp, Christian G. Schanz, Marlene Staginnus, Anja S. Göritz, and Tanja Michael. "Impact of COVID-19 on public mental health and the buffering effect of a sense of coherence." Psychotherapy and Psychosomatics 89, no. 6 (2020): 386-392.

[3] Bish, Connie L., Heidi MichelsBlanck, Mary K. Serdula, Michele Marcus, Harold W. Kohl III, and Laura Kettel Khan. "Diet and physical activity behaviors among Americans trying to lose weight: 2000 Behavioral Risk Factor Surveillance System." Obesity research 13, no. 3 (2005): 596-607.

[4] Centers for Disease Control and Prevention. "Behavioral risk factor surveillance system questionnaire." System 83, no. 12 (2011): 76.

[5] Brooks, Samantha K., Rebecca K. Webster, Louise E. Smith, Lisa Woodland, Simon Wessely, Neil Greenberg, and Gideon James Rubin. "The psychological impact of quarantine and how to reduce it: rapid review of the evidence." The Lancet 395, no. 10227 (2020): 912-920.

[6] Cortez, Pedro Afonso, Shijo John Joseph, Nileswar Das, Samrat Singh Bhandari, and Sheikh Shoib. "Tools to measure the psychological impact of the COVID-19 pandemic: What do we have in the platter?" Asian Journal of Psychiatry 53 (2020): 102371.

[7] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of Artificial Intelligence Research 16 (2002): 321-357.

[8] Ramentol, Enislay, Yailé Caballero, Rafael Bello, and Francisco Herrera. "SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and under sampling for high imbalanced data-sets using SMOTE and rough sets theory." Knowledge and information systems 33, no. 2 (2012): 245-265.

[9] Giraldo-Forero, Andrés Felipe, Jorge Alberto Jaramillo-Garzón, José Francisco Ruiz-Muñoz, and César Germán Castellanos-Domínguez. "Managing imbalanced data sets in multi-label problems: a case study with the SMOTE algorithm." In Iberoamerican Congress on Pattern Recognition, pp. 334-342. Springer, Berlin, Heidelberg, 2013.

[10] Simon, Gregory, Johan Ormel, Michael Von Korff, and William Barlow. "Health care costs associated with depressive and anxiety disorders in primary care." American Journal of Psychiatry 152, no. 3 (1995): 352-357.

[11] Simon, Gregory E., Michael Von Korff, and William Barlow. "Health care costs of primary care patients with recognized depression." Archives of general psychiatry 52, no. 10 (1995): 850-856.

[12] Ciechanowski, Paul S., Wayne J. Katon, and Joan E. Russo. "Depression and diabetes: impact of depressive symptoms on adherence, function, and costs." Archives of internal medicine 160, no. 21 (2000): 3278-3285.

[13] Barua, Sukarna, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. "MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning." IEEE Transactions on knowledge and data engineering 26, no. 2 (2012): 405-425.

[14] Barua, Sukarna, Md Monirul Islam, and Kazuyuki Murase. "A novel synthetic minority oversampling technique for imbalanced data set learning." In International Conference on Neural Information Processing, pp. 735-744. Springer, Berlin, Heidelberg, 2011.

[15] Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." Journal of machine learning research 2, no. Nov (2001): 45-66.

[16] Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic." Journal of Intelligent Learning Systems and Applications 9, no. 01 (2017): 1.

[17] T. Kolenik and M. Gams, "Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?" in IEEE Technology and Society Magazine, vol. 40, no. 1, pp. 80-86, March 2021, doi: 10.1109/MTS.2021.3056288.

[18] Kolenik, T. (2022). Methods in digital mental health: smartphone-based assessment and intervention for stress, anxiety, and depression. In Integrating Artificial Intelligence and IoT for Advanced Health Informatics (pp. 105-128). Springer, Cham.

[19] K. Nigam, K. Godani, D. Sharma, S. Khandelwal and M. Rathi, Personalised Heart Monitoring and Reporting System. 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH), 2020, pp. 68-73, doi: 10.1109/INBUSH46973.2020.9392184.

[20] Rathi, M., Sahu, S., Goel, A., & Gupta, P. (2022). Personalized Health Framework for Visually Impaired. Informatica, 46(1).

[21] Gautam, A., Chauhan, A. S., Srivastava, A., Jadon, C., & Rathi, M. (2019). Major Histocompatibility Complex Binding and Various Health Parameters Analysis. In Smart Healthcare Systems (pp. 151-164). CRC Press.

[22] Rathi, M., Mittal, A., & Agarwal, D. (2020, February). Prediction of Thorax Diseases Using Deep and Transfer Learning. In 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH) (pp. 236-240). IEEE.

[23] Rathi, M., & Pareek, V. (2016). Disease prediction tool: an integrated hybrid data mining approach for healthcare. IRACST Int J Comput Sci Inf Technol Secur (IJCSITS), 6(6), 32-40.

[24] O. Oyebode, F. Alqahtani and R. Orji, "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews," in IEEE Access, vol. 8, pp. 111141-111158, 2020, doi: 10.1109/ACCESS.2020.3002176.

[25] E. Gore and S. Rathi, "Surveying Machine Learning Algorithms On Eeg Signals Data For Mental Health Assessment," 2019 IEEE Pune Section International Conference (PuneCon), 2019, pp. 1-6, doi: 10.1109/PuneCon46936.2019.9105749.

[26] Sabourin, A. A., Prater, J. C., & Mason, N. A. (2019). Assessment of mental health in doctor of pharmacy students. Currents in Pharmacy Teaching and Learning, 11(3), 243-250.

[27] Hou, Y., Xu, J., Huang, Y., & Ma, X. (2016, November). A big data application to predict depression in the university based on the reading habits. In 2016 3rd International Conference on Systems and Informatics (ICSAI) (pp. 1085-1089). IEEE.

[28] Gokten, E. S., & Uyulan, C. (2021). Prediction of the development of depression and post-traumatic stress disorder in sexually abused children using a random forest classifier. Journal of Affective Disorders, 279, 256-265.

[29] Xin, Y., Ren, X. Predicting depression among rural and urban disabled elderly in China using a random forest classifier. BMC Psychiatry 22, 118 (2022). https://doi.org/10.1186/s12888-022-03742-4.

[30] Srividya, M., Mohanavalli, S., & Bhalaji, N. (2018). Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems, 42(5), 1-12.

[31] Tate, A. E., McCabe, R. C., Larsson, H., Lundström, S., Lichtenstein, P., & Kuja-Halkola, R. (2020). Predicting mental health problems in adolescence using machine learning techniques. PloS one, 15(4), e0230389.

[32] Reddy, U. S., Thota, A. V., & Dharun, A. (2018). Machine learning techniques for stress prediction in working employees. In 2018 IEEE International

[33] Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-4). IEEE.

[34] Potter, G., Wong, J., Alcaraz, I., & Chi, P. (2016). Web application teaching tools for statistics using R and shiny. Technology Innovations in Statistics Education, 9(1).

[35] Conway, Jake R., Alexander Lex, and Nils Gehlenborg. "UpSetR: an R package for the visualization of intersecting sets and their properties." Bioinformatics 33, no. 18 (2017): 2938-2940.

[36] Sinha, A., & Rathi, M. (2022). Advanced Computational Techniques for Sustainable Computing. ISBN 9781003046431, Taylor & Francis, CRC Press, Chapman & Hall, pp. 1-338

[37] Adwitiya Sinha, "PSIR: A Novel Phase-wise Diffusion Model for Lockdown Analysis of COVID-19 Pandemic in India," System Assurance Engineering & Management, Springer, pp. 1-17, October 2021

[38] Ramanna, Sheela, and Lakhmi C. Jain. Emerging paradigms in machine learning. Edited by Robert J. Howlett. Heidelberg: Springer, 2013.

[39] Sinha, A., & Rathi, M. (2021). COVID-19 prediction using AI analytics for South Korea. Applied Intelligence, 51(12), 8579-8597.

[40] Sinha, A. (2021). PSIR: a novel phase-wise diffusion model for lockdown analysis of COVID-19 pandemic in India. International Journal of System Assurance Engineering and Management, Springer, 1-14.

[41] Saxena, N., Chahal, E. S., Sinha, A., & Chand, S. (2021). Coronavirus Infection Segmentation & Detection Using UNET Deep Learning Architecture. In 2021 IEEE 18th India Council International Conference (INDICON), pp. 1-6.

[42] Gjoreski, M., Mitrevski, B., Luštrek, M., & Gams, M. (2018). An inter-domain study for arousal recognition from physiological signals. Informatica, 42(1).

[43] Peng, X. (2021). Research on Emotion Recognition Based on Deep Learning for Mental Health. Informatica, 45(1).

[44] Adeniji, O. D., Adeyemi, S. O., & Ajagbe, S. A. (2022). An Improved Bagging Ensemble in Predicting Mental Disorder using Hybridized Random Forest-Artificial Neural Network Model. Informatica, 46(4).

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several sci- entific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute tem- perature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, en- ergy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research de- partments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the uni- versities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; ap- plied mathematics. Most of the activities are more or less closely connected to information sciences, in particu- lar computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automa- tion and control, professional electronics, digital communi- cations and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the in dependent state of S**love**nia (or S nia). The capital today isconsidered a crossroad between East, West and Mediter-ranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strat- egy for technological development to foster synergies be- tween research and industry, to promote joint ventures be- tween university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please register as an author and submit a manuscript at: http://www.informatica.si. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosoph- ical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be pub- lished within one year of receipt of email with the text in Infor- matica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing edi- tor.

SUBSCRIPTION

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twentyeight years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers ac- cepted by at least two referees outside the author's country. In ad- dition, it contains information about conferences, opinions, criti- cal examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and infor- mation industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica web edition is free of charge and accessible at http://www.informatica.si.
Informatica print edition is free of charge for major scientific, ed- ucational and governmental institutions. Others should subscribe.

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2022 (Volume 46) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Blaž Mahnič, Gašper Slapničar; gasper.slapnicar@ijs.si
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Slovene Society for Pattern Recognition (Vitomir Štruc)
Slovenian Artificial Intelligence Society (Sašo Džeroski)
Cognitive Science Society (Olga Markič)
Slovenian Society of Mathematicians, Physicists and Astronomers (Dragan Mihailović)
Automatic Control Society of Slovenia (Giovanni Godena)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Mark Pleško)
ACM Slovenia (Nikolaj Zimic)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 47 Number 1 March 2023 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **47** (2023) Number 1, pp. 1–142