Autonomous Artificial Intelligence Systems for Fraud Detection and Forensics in Dark Web Environments

[1]Romil Rawat, [2]Olukayode A. Oki, [3]Rajesh Kumar Chakrawarti, [4]Temitope Samson Adekunle,
[5]Jose Manappattukunnel Lukose, [6]Sunday Adeola Ajagbe
[1]Department of Computer Architecture and Communications, University of Extremadura, 06006, Badajoz, Spain
[2]Department of Information Technology, East London, Walter Sisulu University, South Africa
[3]Department of Computer Science and Engineering, Sushila Devi Bansal College, Bansal Group of Institutions, Indore, India
[4]Colorado State University
[5]Department of Information Technology, Walter Sisulu University, South Africa.
[6]Department of Computer & Industrial Production Engineering, First Technical University, Ibadan, 200255, Nigeris
E-mail: rawat.romil@gmail.com, ooki@wsu.ac.za, dean.cse@sdbc.ac.in, temitope.adekunle@colostate.edu, joseariasgon6@gmail.com, sunday.ajagbe@tech-u.edu.ng

Keywords: artificial intelligence, crime, forensics, cyber terrorism, cyberattack, IoT

Received: November 28, 2022

Artificial Intelligence (AI) influenced technical aspects of research for generating automated intelligent behaviors covering divergent domains but has shown appreciable results when used in cyber forensic technology for crime analysis and detection. AI experts warned about possible security risks associated with algorithms and training data, as AI inherits computing features dealing with IoT-based Smart applications and autonomous transportation, and may be found to be susceptible to vulnerability and threats. The present work discusses models for analyzing terrorists related information and categorizing malicious events by covering the literature review on security risks and AI-related criminality by presenting a taxonomy of criminal behavior and signatures covering tools and criminal targets used in fraudulent activities using AI features by digital forensic techniques. We've also shown how AI may make existing crimes more potent, and that new sorts of crimes could emerge that haven't been identified previously. This study has presented a systematic structure for AI crime and dealing strategies. Furthermore, we have proposed AI forensics, a unique strategy for combating AI crime. We discovered that several concepts of DF are still not preferred in AI-based forensics after conducting a comparative examination of forensics.

Povzetek: Narejen je pregled sistemov umetne inteligence za ugotavljanje kriminalnih spletnih aktivnosti z dodatno analizo, kje je možno AI uporabiti za te namene.

1 Introduction

Artificial Intelligence (AI) algorithm imitated by human intelligence based on inference approach, and Deep Learning (DL) based on brain morphology has sparked several applications for detecting, predicting, generating models and analyzing the issues relating to Online Social networks (OSN) [1, 2]. Vulnerability, Forensic technology (for Smart device, memory, Cloud) in criminology, IoT and Automations, Healthcare, NLP based on huge analytics of real-time data. The dramatical implementation of DL techniques made the easy access of using online business applications relating to the financial transaction, and these featured processes attracted the cyber threat (phishing, hacking, fraud) originated by cybercriminals, some are operating via the dark web (DW), anonymous porting of hidden internet area, affecting directly and indirectly, economy, security and operation on the internet [3, 4, 5]. Figure 1 shows the fraud framework affecting organizations. This extraordinary rise in AI and DL usage brings to mind the early days of information and communication technology (ICT) for AI stakeholders. Unexpected issues (cybercrime and terrorism, security breach, privacy violation, Extortion, Illicit trade etc.) arose when ICT grew at rapid rates in the past, resulting in significant social value disturbance [6, 7]. Similarly, there are rising concerns regarding the potential for AI to generate a variety of issues [8]. As Brundage & Skarmeta, [9] pointed out, the shifting danger of technological environment up-gradation necessitates significant consideration of research for avoiding and minimizing the dark side of AI [10]. Here, in this section, AI-based security risks, anticipated crimes, and digital forensics (DF) are discussed. The famous DW market silk road operated using the onion router (ToR), an anonymous stream of E-Business and trade, focused on illegal

ERRATA: In the first accepted version of the paper published on October 26, 2023 in the author list was another author: José Luis Arias Gonzáles which turned out to be an error in the submission. This corrected author list was published on Web on November 9, 2023.

activities based on digital currency (Cypto-Coin) for transaction and revenue generation, hidden from the outer world or even unnoticed by forensic experts, cyber policing and security agencies. Terrorist organizations use DW- Online Social Network (OSN) (even Facebook has the DW -Onion Platform version) messaging system (like hidden wiki and DuckDuckGo) for managing to fund, terrorist recruitment, propaganda generation, attack operation order declaration, threatening post sharing and tracing government and military operations and so on [11].

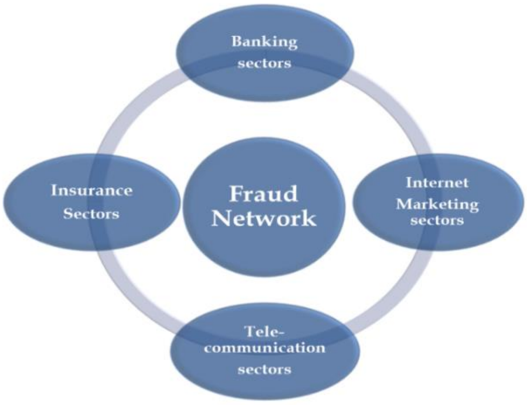


Figure 1: Fraud framework affecting organizations.

For tracing the illicit activities of criminals and terrorists, researchers and security agencies continuous trace OSN for abnormal behavior and post, creating a disturbance and made unauthorized by the government, several models are generated to analyze vulnerable inputs and associated files based on AI and DL techniques [12, 13]. Figure 2 shows the model for tracing and generating a similar dataset comprising fingerprints and signatures of vulnerable events posted on OSN.

a) **Data Collection Program** – Identification of OSN platform, containing malicious and terrorism-related posts, messages, and files.
b) **Terrorism API** – Running Crawlers for collecting details based on a malicious list generated containing hashtag, Keywords, Symbols, Images of weapons and Hidden face-masked Pictures [14].
c) **Dataset** – Organization and identification of severe threatening incidents post for tracking the locations and associated members of groups associated with past (Like, reply, sharing, comments, forwarding).
d) **Classification Program** – Based on the dataset obtained, details are classified for creating types of events and threats planned by cybercriminals with people targeted.
e) **Terrorism data cleaning -** Further threatening task associated with the bombing, massive killing and government attack is classified.
f) **Analysis-** identification for Analyzing the triggering event and timeframe or organized crime for predicting future events with the maximum possibility of new threats and attacks.
g) **Final dataset –** it contains filtered events with parameters and signatures, to be used by security agencies for acting on it.

Rationale

- There is less work available, with a focus on cybercrime forensic evaluation based on Automated techniques for crime detection and prediction.
- The proposed work focused on AI-based automated intelligence behaviour systems for malicious and illicit events.
- The current project focuses on the taxonomy of criminal behaviour and signatures for generating crime recommendations for security officials.
- The work focuses on forensic experts' evaluation of crime behaviours and techniques for the prediction of future illicit crimes.

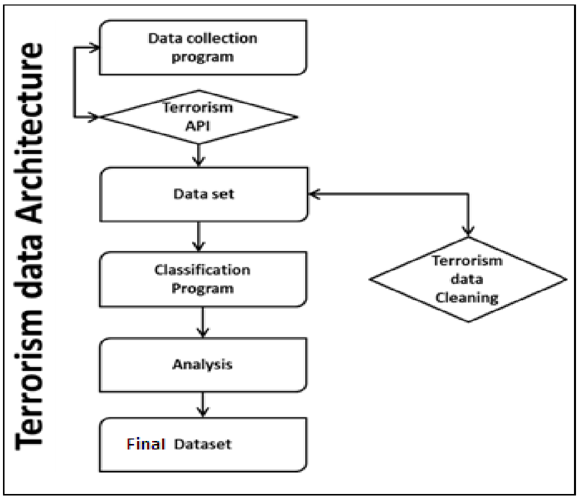


Figure 2: Terrorism activities data architecture

Organization of paper

The remaining parts of the paper are organized as follows: Section 2 shows related work; Section 3 outlines for Use of A.I. - Tool In Crime; Section 4 shows AI -A Target Crime; Section 5 represents the Discussion; Finally, Section 6 concludes the paper with future work.

2 Related work

In this section, we present AI applications with respect to the security threat, criminality, cybersecurity and DF.

a) AI-Based security threats

Researchers coined the phrase 'AI crime,' associated with law, regulations and ethical principles. Several pieces of research have highlighted security concerns and malevolent applications of causing numerous crimes, despite the fact that the phrase AI crime, is still not addressed in the research domain and ICT field [15]. Adopting online identities, known as social bots that behave like humans is a classic example of malevolent AI usage [16]. Socialbot is designed to spread awareness for collaborating people [17] but is used for conducting fraudulent activities at OSN [12]. ML can be designed as weaponizing tool for creating a disturbance at OSN. AI allows huge-forged tweets or post-production, routing towards phishing sites, to be broadcast on any OSN channels causing mass destruction by conveying wrong information [18]. Dangerous social bot activity depends upon user's prior actions and a profile, detecting it has become a computer security problem [19]. When harmful social bots are meant to carry out a political threat, influencing public opinion, damages the personal reputation.

According to some experts including [20], hacker's community had begun to weaponise threatening AI approaches to improve associated cracking abilities invention and newly developed forms of Online threats, used to improve tactics for classic cyber crimes including financial fraud, cyberterrorism, and cyberextortion, among others. When hackers attempt audio phishing, for example, they can mislead victims by imitating the sounds of the victims' relatives or friends realistically [2].

Unlike the previous research, which focused on the issues that certain approaches may create, Brundage et al. [15] provided a holistic view of AI's harmful usage. They focused on three shifts in the danger landscape: the growth of current risks, the development of new threats, and a shift in the threat's usual nature. The cost of jobs that need human labour might be reduced because of the AI system's scalability. As a result of the cost-cutting tactics (e.g. mass spear phishing), perpetrators are able to attack more targets, resulting in the growth of existing risks. New dangers may potentially develop to perform jobs that are impossible for humans to complete (e.g., mimicking people's voices, commanding numerous drones) [21]. The traditional word set (Character Sequence) of vulnerabilities can be altered if highly successful AI threat engines become increasingly frequent. Security domains were also divided into three categories: digital, physical and political [15]. Cyberattacks that target human behavior or AI automated systems are included in the digital security realm. Physical threats, forcing autonomous cars for crashing and manipulating thousands of UAV (unmanned aerial vehicles), are under the physical security realm. Novel dangers in profile degradation and targeted misinformation efforts are all part of the political security area.

By adopting the phrase 'AI crime,' King et al. [22] presented a fresh perspective on AI-based threat concerns and looked at the issue from a different angle. The article divides AI crime into three categories (trade, fiscal market, and bankruptcy) market news manipulation, value price rigging, toxic or hazardous medicines promotion, fraudulent activity, forgery, phishing, credit card fraud, harassment, torture, favour of sexual assault. Each of the offences is classified as containing one or more threats focusing on human behaviors, when defining AI security threats, the psychological threat, and influence by AI Models relating to cognition, a disturbed state of human mind motivating crime conduction [23]. Research focusing on AI threats and privacy concerns arising from the processing of personalized information's relating to e-healthcare, finance, and education may cause security breaches and information disclosure. The data collection (scripts by code designers) and selling is the new business for acquiring confidential records based on AI performance is inherent security risks [24].

GDPR was used to tackle the issue of privacy. The author emphasized that it may be applied to AI when it manages personal data. The preceding research has three consequences for AI stakeholders [25]. First, owing to AI's divergent behaviors among designers should be aware of its exploitation framework to conduct criminal threats, even those intended for legal purposes. Because AI easy to be manipulated approach, anyone involved in the domain must adhere to stringent professional ethics. Second, completely new sorts of security risks will arise that have never been considered before. Because AI can perform activities that were previously thought to be difficult for people or traditional programmes to complete, the risks will be outside the core area of recognized threats. To prevent AI security concerns and respond to AI crime, AI algorithm designers collaborated and work with specialists from many fields. Finally, the AI security field should learn from the cybersecurity industry's mistakes [26]. The predicted AI crimes are extremely intimately engaged in cybercrime, as revealed in earlier research. To prevent AI security concerns and respond to AI crime, AI researchers should work closely with experts from a variety of fields. Finally, the AI security field should learn from the cybersecurity field's mistakes.

b) Digital crime

The evil side of cyberspace is referred to as cybercrime, divided into computer crime as (target and a tool). The goal of digital crime is to disrupt or destroy systems. As a result, cybercriminals involved in activities (terrorism, extortion and warfare,) execute a computer-as-a-targeted host, procedures that have been created to break into computer systems via malware codes (worms, viruses, and spyware) [27]. In the meantime, every data in our daily lives has been digitized, from personal to professional. This shift allows for online severe crimes such as child abuse and stalking, known as computing machines as a tool for crime. Because most threatening tactics are focused on exploiting the weaknesses of potential targets, cybercrime is inextricably linked to cybersecurity. The taxonomy of cybercrime has aided in the development of practical ways to combat the crime. When DF investigators look for crime by focusing on establishing perpetrator's previous conduct for analyzing, any unlawful activity crime scene construction is required. Criminals that utilize computers as tools typically employ well-known tools and exploit infrastructures such as text messages and websites social (OSN) media etc. When investigating a computer as a targeted crime, however, detectives concentrate on harmful applications. To rapidly respond to the crime and determine scopes of harm, they locate the threat (malware) followed by reverse-engineering to determine the virus's goal and the source of triggering [28].

c) Forensics In the digital age

The application of scientifically developed and validated procedures in preservation, gathering, validation, issue identification, data analysis, features interpretation, recording, and digital evidence presentation is defined as DF [20]. Many principles and guidelines have been proposed in the field of DF since each nation and organization has its own set of laws and regulations. Nonetheless, they all share the basic premise that a forensic process is only deemed forensically sound if it adheres to the following five principle components: Meaningful existence of collected evidence (Originality), reporting of generated issues (Errors), Correctness and verified procedures (Transparency and Trustworthiness), quality of collected pieces of evidence (Reproducibility), and case solving expertise (Experience) [15]. The evidence would be difficult to accept in court if the forensic procedure did not follow any of the principles. As a result, investigators must gather and evaluate evidence while following the guidelines. In addition, forensic researchers established a proactive method called Digital Forensic Readiness (DFR) that is intended to handle issues before they happen [29]. During an incident response, DFR attempts to act speedily for gathering and associating digital evidence correctly by reducing the expenses of performing forensic investigation with reducing the risk of losing original assets as a result of a security event. Because the occurrences are caused by information system vulnerabilities, DFR can help prevent or detect cybercrime. Digital forensic experts, like those in other fields, have looked at the use of AI in investigations. To identify malware, Debbabi and Karbab [28] employed an NLP based system with supervised ML. In some datasets, they have an appreciable f1-score (94 %).

Table 1: Related works of overview papers

Ref.	Approach
[30]	DL based forensic knowledge graph method is utilized based on assembles pieces of details from security logs
[31]	Mathematical evidence theory and a probabilistic technique is used for tracing threat Intention Analysis.
[32]	ML based approaches are discussed for cyber forensics
[33]	DF and BD for digital investigators for crime analysis is discussed.
[34]	VSCBR is discussed for assisting investigators predicting evidential Locations.
Proposed Work	The proposed work focused on AI-based automated intelligence and forensic behaviour systems for malicious and illicit events.

3 Use of AI. - tool in crime detection

Given AI's dual-use nature, this section portrays predicted AI as a tool for crime. Because AI systems are built on digital infrastructure, they are vulnerable to cybercrime, which includes both computer-as-tool and computer-as-target crimes. Furthermore, AI may be used to operate autonomous equipment such as smart cars, drones, and Internet of Things (IoT) devices [6] to commit physical crimes [35]. In this part, we'll look at how AI may be utilized to sharpen cyber attacks. Then we turn our attention to physical crime, which is considered a fresh attack.

a) Cybercrime with advanced technology

Perpetrators can use AI approaches to commit innovative cybercrime that was previously thought to be impossible to carry out. This section looks at how AI may be used to combat cybercrime.

i. Crime with a computer as a tool

Previous studies have shown that AI may be utilized in website phishing [8], Scam email with the use of AI in business has been extensively researched. The targeted advertising strategy, based on consumer’s prior purchasing history (interests), used by attackers motivates criminals to generate targeted phishing sites for the click, sometimes containing AI-based chatbot. AI will improve tactics for defrauding clients through the use of a malevolent chatbot [36]. The chatbot may engage with consumers continuously and collect large amounts of data about their behavior and preferences. In academics and industry, the chatbot has already been developed and is in use. Initially, the chatbot was primarily text-based. However, as NLP approaches progressed, the chatbot has been designed to communicate with humans vocally. While some research shows about AI-assisted voice might be useful in relaxation [(social therapy), education, and medical diagnostics, on the other side AI-assisted voice could increase fraudulent activities [36]. Because speech is biometrics, an indispensable feature in security mechanisms, and could be a powerful tool for attackers (like voice phishing).

Fake news (originated by virtual anchors) associated with crime in OSN (like Facebook, Twitter, Instagram and YouTube) gained extreme attention and had a significant impact on political problems (policy decisions, propaganda, extreme view of support or hate in elections) [6, 2]. Fake news gains increasing strength with the deep fake technology. Fake videos imitating renowned politicians can hurt people by spreading incorrect details [37]. News organizations built an AI-based anchor for improving their efficacy and expenses. The projection for the market in 2023 surpasses 62 billion U.S. dollars [21]. Organizations offering strategies to prevent fraudulent insurance claims, data fraud, and money laundering are competing with attacking automation (AI-based designs), Around 24 % of OSN users have been a casualty of online data fraud [38]. Figure 3 shows global statistics on fraudulent activities. These malfeasance exercises can be credit card fraud, or bank fraud having wire transfers representing the most elevated estimation of fraudulent activity for threat originators (Criminals).

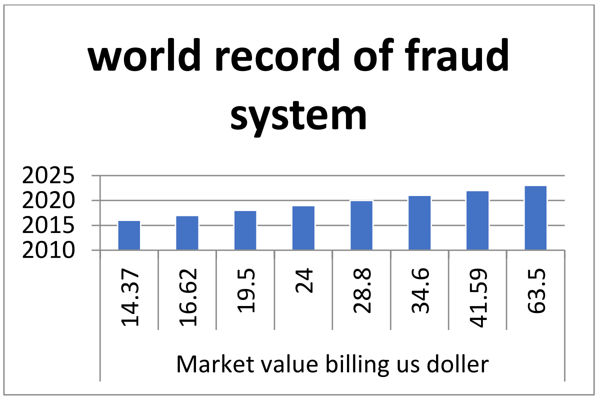


Figure 3: Global statistics of fraudulent activities

ii. The computer as a crime target

AI can accomplish previously unsolvable jobs at a cheaper cost and with less effort. It can have the same impact as hiring extra-human analysts by producing duplicates of the AI system. As a result of this feature, attackers can get illegal access to authentication procedures. A dictionary attack is the most efficient method of obtaining the password because it employs words or phrases (Relating to user attributes) for the password [6]. The social engineering approach of obtaining victim's details from the internet (like birthdates, phone numbers, city, hobby, address, etc.) is frequently utilized while constructing the dictionary. Collecting information takes time and deep effort compilation, but the AI-based system easily and quickly processes the automation for reverse social engineering.

For thieves, an automated detection approach for generating the weaknesses might be a helpful tool. It was suggested that AI might be used to discover vulnerabilities [39]. They showed that using CNN and a tree ensemble (TE) over a standard static analyzer has certain advantages. A technique for detecting network flaws was proposed [40]. Without studying source code, applications with vulnerabilities might be identified using the suggested technique. Aside from those investigations, different approaches for detecting susceptibility are being investigated. Despite the fact that the approaches were created for the general benefit, criminals may utilize them to identify susceptible systems.

b) Physical crime

The issue of AI security goes outside cyberspace, especially with the growing use of IoT [41]. A criminal can physically assault a target by controlling an AI system (vehicle, house, animal, Civilian and, Militant) [42]. The ethics of AI in relation to physical crime has been explored in the field of science ethics represented robotic ethics by stating that AI-based robots could attack and murder people intentionally or unintentionally [43]. AI systems might inflict bodily injury, making responsibility (legal and moral) for the associated effect can be difficult as observed [44]. Military AI – Automotive system, on the other hand, has been created for military purposes and is naturally geared to attack physical targets and towards the benefit of the public, but it may also be utilized to hurt individuals (like Swarm robotics), the following conditions must be satisfied in order to run the drone swarm [45]. Self-governing (not under centralized control).

i. Capable of perceiving their immediate surroundings as well as the presence of other swarm members nearby.
ii. Capable of communicating with others in the swarm on a local level.
iii. Swarming is the ability of a group of computers to work together to complete a job. Traditional programming algorithms struggled to fulfil the criteria, but advances in AI have enabled swarming.

Robotic systems can use this swarming technique in automobiles and are able to remotely modify car vulnerabilities. By exploiting weaknesses in the Sevcon Gen4 controller, an Electric Control Unit (ECU) was fitted in Twizy. Various threat scenarios for Renault Twizy 80, an E- automobile were presented [46]. Despite the fact that the suggested approach is only effective while the automobile is powered on, they demonstrated that after hacking, the attacker may remotely manipulate the vehicle system. The threats found in Controller Area Network (CAN) protocol, which was widely considered to be the industry standard designed for vehicle networking was presented [47]. Using the level of risk, became successful in their message injection attack use for inducing ECU failures [4]. Swarm technology is a major concern because the damage-causing physical harm to the system component.

4 Artificial intelligence -A target crime

AI as a target crime damage or impairs the processing of information or the operation of an AI automated processes. The AI system is made up of two parts (training and inference). Based on the training dataset, the system creates a model. To categorize the data, the trained model is employed in the inference system (IS) [3]. The technology develops an algorithm, that identifies civilians from among Terrorist Groups -Militant. The method is loaded into the inference system, which then decides if the object picture acquired from sensors is correct (either a Civilian or a Militant).

Threat models based on adversarial examples (AE) were proposed in several types of research, concentrated on the consequences of malicious data injection (Poisoning) during the phases (training or inference) [48]. The tests showed that when AI systems are targeted by AEs vulnerability (like malware), their performance suffers. AEs that can avoid detection are used to target systems. The impersonate attack entails feeding the inference system with mimicked data samples that can incorrectly categorize the real samples. The inversion attack is used to deduce specific aspects from the AI system's output towards input. In addition, a complete view of the vulnerability (threat) model and discussed AI trust model, attack surface, adversarial capabilities and purpose with a focus on the models' features (privacy, confidentiality and integrity) was offered [28]. Figure 4 presents the modeling of AI system for objection identification in this paper.

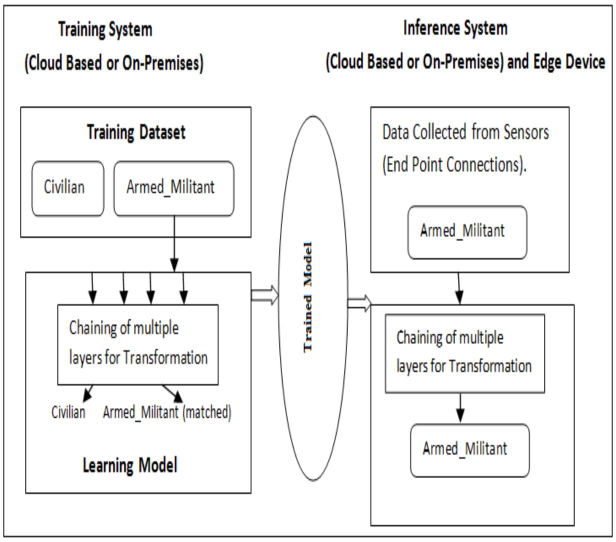

Training System
(Cloud Based or On-Premises)
Inference System
(Cloud Based or On-Premises) and Edge Device
Training Dataset
Civilian
Armed_Militant
Chaining of multiple layers for Transformation
Civilian
Armed_Militant (matched)
Learning Model
Trained Model
Data Collected from Sensors (End Point Connections).
Armed_Militant
Chaining of multiple layers for Transformation
Armed_Militant

Figure 4: Modeling of AI system for Objection identification

a) Target crime: training system (TS)

Insider spying, malicious external storage (MES) and advanced persistent threat (APT), the technique used to protect the system against hacking. The AI system would suffer significant damage if the training system's security was hacked and in particular, contains a training dataset that has a significant impact on the learning model's effectiveness and effects.

i. Attack at training system (TS)

By lowering the AI's probability in crime prediction affects the AI capability to classify fresh data at inference system (IS) by introducing AEs or altering the current dataset and learning algorithm, which is known as logic corruption, the training system will be significantly harmed and classified as (data Poisoning, manipulation, and logic alteration and corruption) [15]. By introducing AEs, the data Poisoning threat affects the behavior of AI systems. AI incorrectly detects a picture of pandas as a gibbon by associating the unnoticeable noise [49].

Purpose of AEs to deceive AI having the lowest amount of disturbance.

$$\overrightarrow{x^*} = \vec{x} + \arg\min\{\vec{z}: \tilde{o}(\vec{x}+\vec{z}) \neq \tilde{o}(\vec{x})\}$$

The original data is x, and the perturbation is z. To make it an AEx*, the noise was applied to the original data. The letter O stands for oracle, which is a system that replies to questions. A query, mostly utilized in the cryptography community [11], the technique for creating and using AEs has changed and is extensively researched, particularly in the area of image recognition. By the multi-dimensional library of phenotypic elites, Nguyen et al. [41] presented an approach for producing AEs that are completely unidentifiable to human sight (MAP-Elites). Furthermore, AEs may be used in physical space (self-driving automobiles) by presenting a potential threat in wrongly analyzing road signs. AEs may also be created and used for disrupting malware analysis prediction and detection and intrusion detection.

The data alteration crime occurs when offenders have the authorization to change or remove some training data, allowing them to carry out deadly assaults on AI systems. Altering the labels of certain training data, demonstrated the label contamination attack (LCA) impairing AI performance [15]. The most significant crime in the training system is logic corruption. By interfering with the learning algorithm, thieves can change the architecture and parameters of the learned model [19]. If the CNN framework is hacked and corrupted, intruders have control over each of the layers (input, classification, and training parameters).

ii. Theft of training system (TS)

TS consist of three components (training dataset, learning model, and learned model). The training method is considered a trade secret by AI developers and producers of AI-related goods since it is directly connected to the performance of AI. For AI stakeholders, the dataset is critical [4]. They acquire dataset from a variety of sources (like driving-related data, Traffic data and object data) and is a favourite target for criminals taking sensitive data (medical picture, a facial image, or a voice) showing significant privacy violations.

b) Target crime: inference system (IS)

In comparison to threats to the training system, attackers have a relatively easy time accessing the inference system because it is typically installed at end processing devices.

i. Cracking of inference system (IS)

In an inference system, attributes determined during the training phase are the essential components. Depending on where the parameters are located, there are two sorts of operating methods: centralized and distributed models. In the centralized model, the inference procedure is performed by a centralized server created by an AI design provider [3]. The job of the end processing devices (Like IoT devices, smartphones, and in-vehicle entertainment) in a centralized face recognition system (processing of image feature). Centralized approach is theoretically ideal for AI service management in security processing, it may be less beneficial in practice due to the possibility of a bottleneck. As a result, the distributed model is becoming increasingly popular in practical fields.

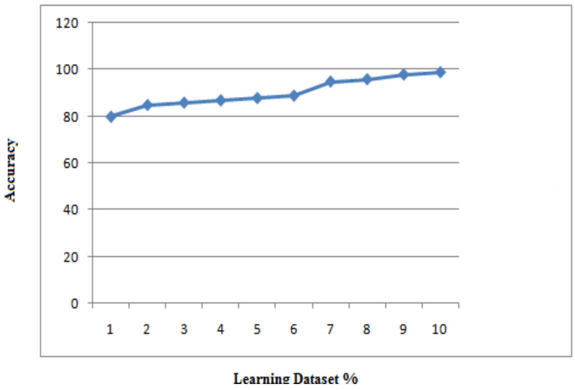
c) AI forensics

To determine 5W1H (when and, where the crime incident is done, who is the criminal, what is the targeted, why the crime commits, and how the vulnerability occurred), forensic investigators should gather and evaluate pieces of evidence based on the operating device platform relating with forensic domains (smartphone, cloud, and IoT). Future AI forensics research directions are presented here to examine the AI technological crimes, the threat detection attributes are presented based on characteristics of AI systems and tactics utilized in forensic for tracing perpetrating activity for AI crimes (Exploration of AI, similarity analysis, detection of adversarial attacks, and damage assessment).

i. Exploration of AI

The models (dataset, learning, trained and, inference), with the application of the AI system used for committing the crime, were collected and analyzed by the investigators to understand the AI's goal based on the assessment having difficulty distinguishing developer's purpose and the AI output. Data and programmes are processed to produce output in traditional programming and utilized to develop an AI model employing random weights among the learning phase, AI parameters are frequently chosen with some unpredictability. As a consequence, even if the same dataset and learning model is used, programmes with distinct parameters and results may be created. i7-8700 CPU and an Nvidia GeForce 1070 Ti graphics card were used in the experiment with variant models to classify binary files as either malicious or benign and models are trained based on the amount of dataset, containing the fraction of data collection. Dataset consisted of 1,500 PE Files for each category. The 800 Malware files were gathered from the Virus Share repository (Open access) [23]. The 800 benign files were gathered from Software Informer, the most reliable source of benign files, as well as system folders established when Windows 10 was first installed. An ensemble approach (voting-based) was employed for increasing the model's performance, using Convolution Neural Networks (CNN). We randomly chose 70% of the dataset 10 times to find performance variations in dataset selection using trained models. It is demonstrated that accuracy varies based on the data used for training. Figure 5 shows the accuracy comparison of trained models.

The findings of the experiments demonstrated that reproducing the AI-based system having less evidence is impossible. Several AI systems employ transfer-based learning, which starts with a pre-trained model, acquiring source data will grow increasingly difficult for technological and policy solutions.



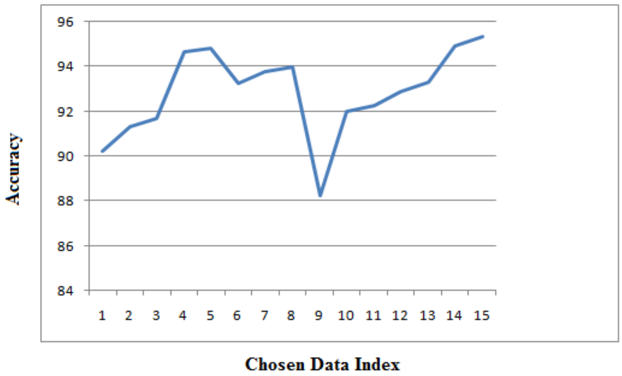


Figure 5: Accuracy comparison of trained models.

ii. Similarity analysis

A classic field of digital forensics involves investigating breaches (copyright infringement, leaks of confidential documents, and privacy infringement). In these instances, similarity analysis (like code plagiarism detection, document similarity, and digital picture similarity) is one of the most significant approaches for detecting illegal activity. Similarity analysis could be utilized to compare datasets (authentic with a suspect) containing (image, audio, text, video) shows the unusual result when having the model of specific datasets. The most crucial area of investigation is the designing of file formats for storing trained models. For example, AI development platforms Python tools (Keras and PyTorch) store and maintain the trained model parameters as an HDF5 file (binary data format) need to be more investigated in the forensics field with model similarity comparison for determining infringement cases, but existing similarity techniques cannot be used on trained models. The trained models are saved as HDF5 files and then use the *ssdeep* and *sdhash* programmes in the digital forensics field to determine the likelihood of similarity between the files [50]. The algorithms decide that the models are not comparable and demonstrate the limitations of existing techniques by suggesting the development of the newer model for calculating similarity.

d) Detection of differential attack

It is complex to proactively prevent (identify) adversarial attacks. Many researchers suggested a defence mechanism to accurately categorize AEs. Due to the difficulty of defending against the adversarial threat, current research has focused on detecting AEs [15, 51]. The hypothesis testings, principle component analysis (PCA) and Bayesian uncertainty estimates (BUE) are examples of statistical approaches to the problem with other available methods of the extraneural network (NN) [15]. Several approaches have been created to identify adversarial threats but AEs threats were devised to neutralize the detection techniques. It is essentially difficult to avoid existing and prospective vulnerabilities since neural network (NN) based classifiers have intrinsic threats leading toward misclassification. As a result, the creation of AEs design is difficult and time-consuming. During incremental training (adapting toward new data for improvement over time.), however, adversarial threats are still a possibility. Intruders can inject AEs into AI-based systems by misrepresenting them as Benign.

e. Assessment of damage

Forensic investigators should determine the degree of harm produced by AI crime. In the case of an AE-based attack, the investigators must determine which details are AEs, how many AEs were really poisoned, and how much the threat changed the situation. Locating AEs is the process of identifying details that increases the prediction rate of error. DNN is used to describe the procedure using the hierarchical composition of m parametric functions $\mathrm{g_i}$. Each $\mathrm{g_i}$ for $\mathrm{j} \in 1 \ldots . \mathrm{m}$ modelled by neurons layer, parameterized by a weight vector θ_i. A DNN model G, computed as follow:

$$G(\vec{y} = g_m(\theta_n, g_{m-1}(\theta_{m-1}, \ldots\ldots g_2(\theta_2, g(\theta_1, \vec{y}))))$$

Assuming that AEs have previously been injected into the target dataset and that the investigators are familiar with training G and the dataset of input-output pairs $(\vec{y}, \vec{z})$, the AEs may be identified using the backward elimination technique as described below:

$$l^* = arg\ min\,\{\textstyle\sum_{\neg(k=l)}^{m} \mathrm{Gj}\ (\overrightarrow{yj})\text{-}z_j\ |\}$$

The L might be a single sample or a group of samples. $\mathrm{G_j}$ stands for trained DNN model without $\mathrm{y_j}$, $\mathrm{g_j}$, or θ_j. The number of misclassifications is used to compute the prediction error in this case. However, because it requires calculating equations (2) and (3) m times, the technique is difficult to use in an actual forensic. B the AI system that includes DNN contains a significant quantity of training data, and calculating the influence of each sample is nearly difficult. The investigators may also be unable to get information on every sample's AI model, creating complexity. As a result, identifying AEs and calculating damage with minimal understanding of the AI model is a significant issue.

5 Discussion

This section emphasizes concerns about AI forensics by comparison with traditional forensics. The current study examines the literature on potential threats and intelligence-related criminal behaviour by presenting a typology of illegal activity and signatures covering tools and criminal targets used in fraudulent activities utilising AI characteristics via digital forensic methods. It also discusses models for categorising malevolent occurrences and analysing terrorist-related details. In order to tackle AI criminality, we propose AI forensics, a novel approach. We conducted a comprehensive analysis of forensics and found that several DF principles are still not accepted in AI-based forensics.

a) Large-scale

Creating an AI model necessitates a large amount of data and resources, which would have been unthinkable in the past. Huge Datastore makes complex for investigators utilizing standard computing-based forensic methods to find data. the large-scale associated issue had been addressed in traditional digital forensics, but AI forensics should encompass a far higher number of data sets. Current AIs, in particular, are mostly focused on multimedia data such as images and sounds and are still complex in digital forensics.

b) Irreproducibility

The inherent unpredictability of AI systems would have an impact on forensic concepts. Several AI approaches employ random features and fail to meet the forensic principles repeatability requirements. If the model of reproducibility test is used solely for AI criminal evidence, like copycat models and AEs, the evidence may

S/N.	AI-based crime modelling	Methods	AI Associated challenges toward Forensics technology
1	Threat originator and Triggering host	AI chatbot, Automated, reverse threat engineering system, Deep Fake, etc.	Exploration of AI
2	Honeypot stations	OSN Vulnerability for Social Engineering threat, Vulnerability scanner, etc.	
3	Hardware (Chip)Module Altering threat	Drone swarm, Chip Alteration, hardware attack, Control management unit modification etc.	
4	Modelling threat for Training System	Vulnerability injection for logic up-gradation, reverse logic triggering	Exploration of AI, Detection of Differential Attack, Detection, Assessment of damage
5	Training system theft	Malware, threat susceptibility-based post, data, Victim system search	Exploration of AI. Similarity Analysis
6	Modelling threat for Inference system		
7	Traditional Forensics	Manually disturbance of records damage of hard disk, Floppy, pan drive, or complete system	Difficult to reproduce
8	AI Forensics	Injecting malicious codes or vulnerability to automatically delete all files or creates encryption and lock.	High-level reverse engineering tools are required for recovering the data

Table 1: AI-Based crimes and Forensics Approach

be dismissed in court, since it does not recreate the scenario at the time of occurrence. However, implementing the reproducibility principles must be examined resulting in additional difficulties, such as arresting the incorrect person. As a result, forensic examiners, legislators, and AI specialists should explore a balance between the stringent and tolerant application of the Models (principles).

c) Expertise

Finally, forensic stakeholders must improve their AI knowledge by having a thorough grasp of AI criminological and forensic methodologies to solve the problems of AI forensics. Traditional available programming (like memory structure, type and compiler) should be understood by forensic investigators. When analyzing malware (in assembly language), they must have prior knowledge of the AI system structure with processing environment for solving forensic issues.

6 Conclusion and future work

Artificial intelligence (AI) is employed in a variety of systems and applications with rising fears that AI might be detrimental to humanity because of its dual-use nature. Intruders may utilize AI engines by abusing the target AI system's inherent weaknesses to carry out illicit actions. This research looked into potential AI-related crimes. We discovered that prior research focuses on malevolent applications of AI system designs in current criminal tactics or the weaknesses of AI approaches and training datasets based on prior art of AI security concerns with the associated crime. We've also shown how AI may make existing crimes more potent, and that new sorts of crimes could emerge that haven't been identified previously. This study has presented a systematic structure for AI crime and dealing strategies.

Furthermore, we have proposed AI forensics, a unique strategy for combating AI crime. We discovered that several concepts of DF are still not preferred in AI-based forensics after conducting a comparative examination of forensics (AI and conventional). Future studies have been recommended to help forensic researchers better grasp the obstacles they confront.

Statement for Conflict of Interest

There are no conflicts to declare.

Author's Contributions: Conceptualization, Methodology, writing of original draft, and Coding/ software **RR**[1], Funding acquisition/APC, Editing, and review, Supervision, and Methodology **OAO**[2], Resources and Writhing of original draft **RKC**[3] **ABO**[4], Literature search and Writing of original draft **JLAG**[5], Conceptualization, and Project administration and supervision **SAA**[6,7]

Informed consent statement

None

Funding "This research received no external funding". "The APC was funded by the Walter Sisulu University, South Africa"

Abbreviations sed

AI	Artificial Intelligence
OSN	Online Social Networks
DL	Deep Learning
NLP	Natural Language Processing
BD	Big Data
DW	Dark Web
ICT	Information And Communication Technology
VSCBR	Variable Scale Case-Based Reasoning Method
ToR	The Onion Router
DF	Digital Forensics
UAV	Unmanned Aerial Vehicles
GDPR	General Data Protection Regulation
DFR	Digital Forensic Readiness
IoT	Internet Of Things
TE	Tree Ensemble
CNN	Convolution Neural Networks
ECU	Electric Control Unit
CAN	Controller Area Network
AE	Adversarial-Examples
IS	Inference-System
MES	Malicious-External-Storage
APT	Advanced-Persistent-Threat
TS	Training System
LCA	Label Contamination Attack
PCA	Principle Component Analysis
BUE	Bayesian Uncertainty Estimates
NN	Neural Network
DNN	Deep Neural Network

References

[1] D. Jeong, "Artificial Intelligence Security Threat, Crime, and Forensics: Taxonomy and Open Issues," *IEEE Access*, vol. 8, pp. 184560-184574, 2020.

[2] P. Sharma, U. Siddanagaiah and G. & Kul, "Towards an AI-Based After-Collision Forensic Analysis Protocol for Autonomous Vehicles," in *2020 IEEE Security and Privacy Workshops (SPW)*, 2020.

[3] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1197-1227, 2015.

[4] B. Hyman, Z. Alisha and S. Gordon, "Secure controls for smart cities; applications in intelligent transportation systems and smart buildings," *International Journal of Science and Engineering Applications*, vol. 8, no. 6, pp. 167-171, 2019.

[5] S. A. Ajagbe, M. O. Oyediran, A. Nayyar, J. A. Awokola and J. F. Al-Amri, "P-acohoneybee: a novel load balancer for cloud computing using mathematical approach," *Computers, Materials & Continua,* vol. 73, no. 1, pp. 1943-1959, 2022.

[6] K. J. Hayward and M. M. Maas, "Artificial intelligence and crime: A primer for criminologists," *Crime, Media, Culture, 1741659020917434.,* 2020.

[7] S. A. Ajagbe, O. A. Oki, O. M. A. and A. Nwanakwaugwu, "Investigating the Efficiency of Deep Learning Models in Bioinspired Object Detection," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Prague, Czech Republic., 2022.

[8] R. Rawat and M. Zodape, "URLAD (URL attack detection)-using SVM," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 1, 2012.

[9] J. B. Bernabe and A. Skarmeta, "Introducing the Challenges in Cybersecurity and Privacy-The European Research Landscape," in *Challenges in Cybersecurity and Privacy—The European Research Landscape*, Rever, 2019, pp. 1-21.

[10] N. Koroniotis, N. Moustafa and E. Sitnikova, "Forensics and deep learning mechanisms for botnets in internet of things: A survey of challenges and solutions," *IEEE Access,* vol. 7, pp. 61764-61785, 2019.

[11] R. Rawat, N. Patearia and S. Dhariwal, "Key Generator based secured system against SQL-Injection attack," *International Journal of Advanced Research in Computer Scienc,* vol. 2, no. 5, 2011.

[12] R. Montasari and R. Hill, "Next-generation digital forensics: Challenges and future paradigms," in *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*, 2019.

[13] I. Bamimore and S. A. Ajagbe, "Design and implementation of smart home nodes for security using radio frequency modules," *International Journal of Digital Signals and Smart Systems,* vol. 4, no. 4, pp. 286-303, 2020.

[14] D. Nahmias, A. Cohen, N. Nissim and Y. Elovici, "Deep feature transfer learning for trusted and automated malware signature generation in private cloud environments," *Neural Networks,* vol. 124, pp. 243-257, 2020.

[15] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel and D. Amodei, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228.,* 2018.

[16] I. Chomiak-Orsa, A. Rot and B. Blaicke, "Artificial Intelligence in Cybersecurity: The Use of AI Along the Cyber Kill Chain," in *12. Chomiak-Orsa, I., Rot, A., & Blaicke, B. (2019, September). Artificial Intelligence in Cybersecurity: The Use of AI Along the Cyber Kill Chain. In International Conference on Computational Collective Intelligence*, 2019.

[17] M. T. Oladejo and L. Jack, "Fraud prevention and detection in a blockchain technology environment: challenges posed to forensic accountants," *International Journal of Economics and Accounting,* vol. 9, no. 4, pp. 315-335, 2020.

[18] R. Rawat, B. Garg, K. Pachlasiya, V. Mahor, S. Telang, M. Chouhan and R. Mishra, "SCNTA: monitoring of network availability and activity for identification of anomalies using machine learning approaches," *International Journal of of Information Technology and Web Engineering (IJITWE),* vol. 17, no. 1, pp. 1-19, 2022.

[19] R. Rawat, V. Mahor, S. Chirgaiya, R. N. Shaw and A. Ghosh, "Sentiment Analysis at Online Social Network for Cyber-Malicious Post Reviews Using Machine Learning Techniques," *Computationally Intelligent Systems and their Applications,* pp. 113-130, 2021.

[20] K. Pipyros, L. Mitrou, D. Gritzalis and T. Apostolopoulos, "A cyber attack evaluation methodology," in *In Proceedings of the 13th European Conference on Cyber Warfare and Security*, 2014.

[21] E. Mantas and C. Patsakis, "Who Watches the New Watchmen? The Challenges for Drone Digital Forensics Investigations," *arXiv preprint arXiv:2105.10917*, 2021.

[22] T. C. King, N. Aggarwal, M. Taddeo and L. Floridi, "Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *Science and engineering ethics,* vol. 26, no. 1, pp. 89-120, 2020.

[23] R. Rawat, V. Mahor, S. Chirgaiya and A. S. Rathore, "Applications of Social Network Analysis to Managing the Investigation of Suspicious Activities in Social Media Platforms," in *Advances in Cybersecurity Management* , Springer, Cham, 2021, pp. 315-335.

[24] L. Zhang, W. A. N. G. Qing and T. I. A. N. Bin, "Security threats and measures for the cyber-physical systems," *The Journal of China Universities of Posts and Telecommunications,* vol. 20, pp. 25-29, 2013.

[25] M. Karyda and L. Mitrou, "Internet forensics: Legal and technical issues," in *Second International Workshop on Digital Forensics and Incident Analysis (WDFIA 2007)* , 2007.

[26] S. Samtani, M. Kantarcioglu and H. Chen, "Trailblazing the artificial intelligence for cybersecurity discipline: a multi-disciplinary research roadmap.," *23. Samtani, S., Kantarcioglu, M., & Chen, H. (2020). Trailblazing the artificial intelligence for cybersecurity discipline: a multi-disciplinary research roadmap.,* 2020.

[27] R. Rawat, A. S. Rajawat, V. Mahor, R. N. Shaw and A. Ghosh, "Dark Web—Onion Hidden Service

Discovery and Crawling for Profiling Morphing, Unstructured Crime and Vulnerabilities Prediction," in *Innovations in Electrical and Electronic Engineering* , Springer, Singapore, 2021, pp. 717-734.

[28] S. Papastergiou, H. Mouratidis and E. M. Kalogeraki, "Cyber security incident handling, warning and response system for the european critical information infrastructures (cybersane)," in *International Conference on Engineering Applications of Neural Networks* , 2019.

[29] M. Elyas, A. Ahmad, S. B. Maynard and A. Lonie, "Digital forensic readiness: Expert perspectives on a theoretical framework," *Computers & Security,* vol. 52, pp. 70-89, 2015.

[30] I. A. K. Tuhin, P. Loh and Z. Wang, "Smart Cybercrime Classification for Digital Forensics with Small Datasets," in *Cyber Security, Cryptology, and Machine Learning: 6th International Symposium, CSCML 2022*, Be'er Sheva, Israel, 2022.

[31] M. Rasmi and K. E. Al-Qawasmi, "Improving Analysis Phase in Network Forensics By Using Attack Intention Analysis," *International Journal of Security and Its Applications,* vol. 10, no. 5, pp. 297-308, 2016.

[32] K. Rajendiran, K. Kannan and Y. Yu, "Applications of machine learning in cyber forensics.," in *Confluence of AI, Machine, and Deep Learning in Cyber Forensics* , IGI Global, 2021, pp. 29-46.

[33] E. E. D. Hemdan and D. H. Manjaiah, "Digital investigation of cybercrimes based on big data analytics using deep learning," in *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications* , IGI Global, 2020, pp. 615-632.

[34] A. Wang and X. Gao, "A variable scale case-based reasoning method for evidence location in digital forensics," *Future Generation Computer Systems,* vol. 221, pp. 209-219, 2021.

[35] B. K. Mohanta, D. Jena, U. Satapathy and S. Patnaik, "Survey on IoT security: Challenges and solution using machine learning, artificial intelligence and blockchain technology," *Internet of Things,* vol. 11, p. 100227, 2020.

[36] J. Kietzmann, J. Paschen and E. Treen, "Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey," *29. Kietzmann, J., Paschen, J., & Treen, E. (2018). Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey. Journal of Advertising Research,* vol. 58, no. 3, pp. 263-267, 2018.

[37] R. Chesney and D. Citron, "Deep fakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Affairs,* vol. 98, p. 147, 2019.

[38] S. Dhariwal, R. Rawat and N. Patearia, "C-Queued Technique against SQL injection attack," *International Journal of Advanced Research in Computer Science,* vol. 5, no. 2, 2011.

[39] A. S. Rajawat, R. Rawat, K. Barhanpurkar, R. N. Shaw and A. Ghosh, "Vulnerability Analysis at Industrial Internet of Things Platform on Dark Web Network Using Computational Intelligence," *Computationally Intelligent Systems and their Applications,* pp. 39-51, 2021.

[40] N. Kaloudi and J. Li, "The ai-based cyber threat landscape: A survey.," *ACM Computing Surveys (CSUR),* vol. 53, no. 1, pp. 1-34, 2020.

[41] B. Mittelstadt, C. Russell and S. Wachter, "Explaining explanations in AI.," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019.

[42] R. B. M. J. &. B. C. Zeid, "Zeid, R. B.; Moubarak, J.; Bassil, C.," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020.

[43] Y. Y. Ke, T. T. Peng, T. K. Yeh, W. Z. Huang, S. E. Chang, S. H. Wu and C. T. ... Chen, "Artificial intelligence approach fighting COVID-19 with repurposing drugs," *Biomedical Journal,* vol. 43, no. 4, pp. 355-362, 2020.

[44] M. U. Scherer, "Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies," *Harv. JL & Technology,* vol. 29, p. 353, 2015.

[45] P. Svenmarck, L. Luotsinen, M. Nilsson and J. Schubert, "Possibilities and challenges for artificial intelligence in military applications," in *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' meeting*, Neuilly-sur-Seine France, 2018.

[46] S. Jafarnejad, L. Codeca, W. Bronzi, R. Frank and T. Engel, "December). A car hacking experiment: When connectivity meets vulnerability," in *2015 IEEE globecom workshops (GC Wkshps)*, 2015.

[47] F. Martinelli, F. Mercaldo, V. Nardone and A. Santone, "Car hacking identification through fuzzy logic algorithms," in *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, 2017.

[48] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li and A. .. K. T. Rahmati, "Physical adversarial examples for object detectors," in *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18).*, 2018.

[49] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572,* 2014.

[50] E. Raff and C. Nicholas, "Lempel-Ziv Jaccard Distance, an effective alternative to ssdeep and sdhash," *Digital Investigation,* vol. 24, pp. 34-49, 2018.

[51] S. A. Ajagbe and M. O. Adigun, "Deep learning techniques for detection and prediction of pandemic

diseases: a systematic literature review," *Multimedia Tools Application*, vol. 2023, pp. 1-35, 2023.