Volume 44 Number 4 December 2020 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.

# Decision Tree for Classification and Regression: A State-of-the Art Review

Monalisa Jena and Satchidananda Dehuri
P.G. Department of Information and Communication Technology
Fakir Mohan University, Balasore, Odisha, India
E-mail: bmonalisa.26@gmail.com, satchi.lapa@gmail.com

**Overview paper**

**Keywords:** data mining, classification, regression, decision tree, prediction

**Received:** December 8, 2019

*Classification and regression are defined under the umbrella of the prediction task of data mining. Discrete values are predicted using classification techniques, whereas regression techniques are most suitable for predicting continuous values. Analysts from different research areas like data mining, statistics, machine learning, pattern recognition, and big data analytics preferred decision trees over other classifiers as it is simple, effective, efficient, and its performance is competitive with others in a few cases. In this paper, we have extensively reviewed many popularly used state-of-the-art decision tree-based techniques for classification. Additionally, this work also reviews some of the decision tree based techniques for regression. We have presented a review of more than forty years of research that has been emphasized on the application of decision tree in both classification and regression. This review could be a potential resource for all the researchers who are keenly interested to apply the decision tree based classification/regression in their research work.*

*Povzetek: V preglednem članku je podana analiza raznovrstnih metod in tehnik odločitvenih in regresijskih dreves za namene rudarjanja podatkov.*

## 1 Introduction

With the advancement of technologies, the process of data generation and collection is increasing at an exponential rate. The embedded sensors, IoTs, ubiquitous devices like scanners, bar code readers, and smartphones generate a huge amount of data at an exponential rate, which contributes to the expansion of data size and volume [1] [2] [3]. Intuitively, the valuable hidden knowledge and information in this huge amount of accumulated data could be the potential source to enhance the decision-making capability of the decision-makers of an organization or society [4] [5] [6]. Some of the classification techniques like decision tree (DT), support vector machine (SVM), and random forest [7] [8] have been proven to be effective models for extracting knowledge, that is valid, potential, novel, and finally useful. In a decision tree, interpretable rules together with the constraints can be extracted by the decision-maker without compromising the performance of the model [9] [10]. A decision tree is an acyclic graphical structure $G(V, E)$, where, $V \in \{V_1, V_2\}$ represents a finite, non-empty set of nodes; $V_1$ represents a set of leaf nodes containing the class values and $V_2$ is the set of intermediate nodes corresponding to one of the attributes. Similarly, the set of edges, $E$ represents distinct attribute values. DT is one of the popularly used classifiers because of its intelligible nature that takes after the human thinking [11]. DT induction algorithms are preferred over other learning algorithms due to their flexibility, robustness to noise, the low computational cost for model construction, and the ability to handle redundant attributes. They are quite simple and easy to understand by human beings and their performance is comparable with others [12] in certain cases. Decision trees can handle both classification and regression tasks. In classification, a discrete value is predicted, whereas a continuous value is predicted through regression [13]. Decision trees are also competent in handling unseen samples having multiple class labels [14].

A sample DT is depicted in Figure 1. In this Figure, the DT is used to identify the types of contact lenses suitable for an individual having a set of features. It employs the *lenses* data set, one of the popular datasets collected from the University of California, Irvine (UCI) Machine Learning repository [15]. In Figure 1, the internal nodes and class labels are represented in the form of ovals and rectangles, respectively. Four different features such as tear production rate, age, spectacle prescription, astigmatic and three class labels namely hard, soft, and none are considered in this example. A path $\{v_1, v_2, ...., v_n\}$ drawn from $v_1$ to $v_n$ represents the class prediction for a tuple, where $v_1$ is the root node, $v_2$ to $v_{n-1}$ are the intermediate nodes, and $v_n$ is the leaf node of that particular path. For example, for the tuple (age: presbyopic, astigmatic: yes, spectacle prescription: myope, tear production rate: reduced), the class label is "none". In this way, several rules can be extracted from the decision tree and using those rules, the class label of

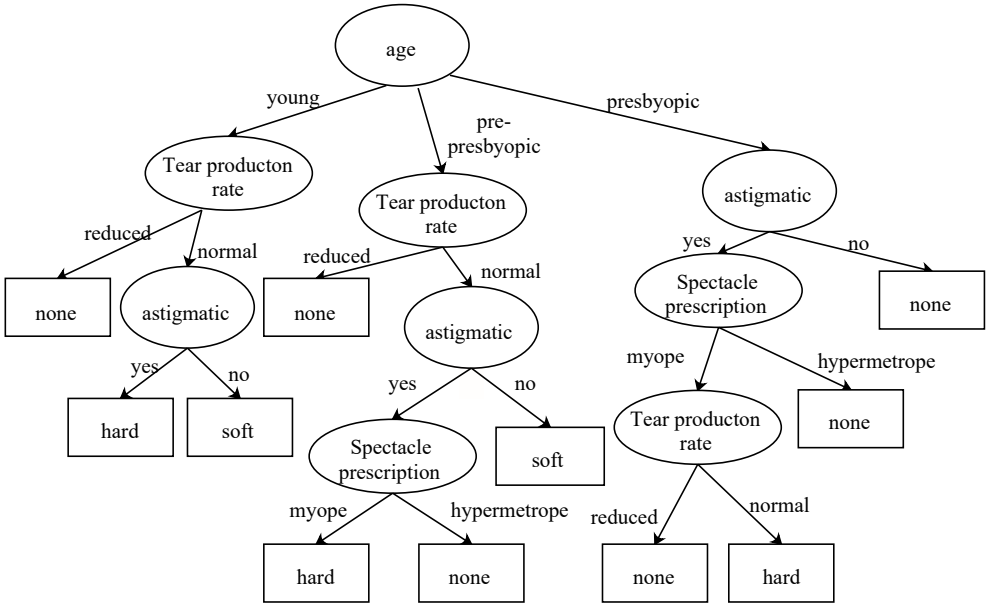



Figure 1: Decision tree to ascertain type of contact lenses to be used by a person.

an unseen sample can be predicted [16]. In Figure 1, the attribute "age" is taken as the root node. The root is selected using several attribute selection measures [17], and the splitting attribute is chosen at a particular node as per the well-defined splitting criterion. For example, the DT in Figure 1 is generated by applying entropy as the attribute selection measure. Hence, age becomes the root node as it is selected as the splitting attribute.

In the past few decades, a number of classification as well as regression tree algorithms have been proposed by several pioneers. Figure 2 gives an overall idea of the number of research papers published in the domain of DT for classification and regression from 1971 to date. An effort has been made to make an extensive review of the different classification and regression tree techniques which would be helpful for the beginners and enthusiastic researchers in this specific field of research. From Figure 2, it can be observed that over the years, research in this particular field has increased spectacularly because of the efficiency, performance, and effectiveness of DT in several application domains. Many researchers in the literature have presented reviews on the classification and regression tree algorithms. Some of the works have missed few parameters while some of them have provided just a brief overview, and some are outdated. Even though we intend to give a balanced discussion, some of the remarks certainly reflect the viewpoints of the authors.

Lim et al. [18], have compared twenty-two decision tree algorithms based on performance parameters like accuracy and computation speed. Classification accuracy is measured by the mean error rate and the mean rank of error rate. Along with the decision tree algorithms, they have also presented nine statistical and two neural network algorithms. They have experimented on these algorithms using thirty-two datasets, out of which fourteen are from real life domains, five are from the STATLOG project, two are synthetic, and the rest are from the UCI repository. Among the decision tree algorithms, QUEST with linear splits is found to have the highest accuracy, and logistic regression is the second best among the thirty three statistical algorithms. Podgorelec et al. [19], have limited their review work on decision trees specific to the field of medicine. They have presented alternatives to the few traditional induction approaches while emphasizing the existing and future applications of medicine. Perlich et al. [20], came up with a large scale comparison between two famous classification models of that time, tree induction and logistic regression. Based on the class membership probabilities, they had estimated classification accuracy and quality of rankings. They have observed that logistic regression performed well for smaller training sets while tree induction methods for comparatively larger datasets.

Rokach and Maimon [21] have presented an updated survey on the induction of decision tree algorithms of that time in a top-down manner. Besides, they suggested a unified algorithmic framework for presenting the decision tree induction algorithms and provided profound descriptions of the various pruning technologies and splitting criteria. They have observed that most of the algorithms fitted the framework with different stopping criteria and pruning methods. Barros et al. [22], have provided a review, which mainly focused on decision tree and evolutionary algorithms. They have presented a taxonomy that designs the decision tree components using evolutionary algorithms. They have also discussed various applications of evolutionary algorithms on decision tree induction in several domains. Loh [23] has presented a brief review of both classification and regression tree algorithms. In his paper, a brief comparison of the classification tree algorithms C4.5, RPART, QUEST, CRUISE, and GUIDE is presented using

prediction accuracy as the performance measure. The author has applied these algorithms on cars dataset for the 1993 model year, and GUIDE appeared to have the highest prediction accuracy. For comparing regression tree models, he has collected data from 654 children aged between 3 and 19 and applied those models on these datasets. GUIDE linear regression tree model was found to have higher prediction accuracy than piecewise constant models. For classification trees, prediction error was measured by misclassification cost, and in the case of regression trees, it was measured by the squared difference between predicted and actual values. Loh [24] again performed a comprehensive review on classification and regression tree algorithms which have been adopted in the last fifty years. In his paper, he focused on the majority of the algorithms that performed consistently well for a long period and for which software was widely available. The review work also provided the developments and key ideas supporting these algorithms. He has also presented a comparative analysis of the classification tree models and their partitions given by all the classification tree models using iris data from the UCI repository. A Similar procedure has been followed for regression tree models using baseball data from Statlib.

In contrast to others, we have presented a survey of all the classification and regression tree algorithms in a technical yet easy to understand manner. We have provided an extensive review of DT algorithms that have consistently better performance and stood the test of the time in the last forty years. We have also discussed the application details of the techniques in various domains under DT for classification as well as regression. This paper would be a potential resource for future researchers and enthusiast readers to get an overall idea about which algorithm works best in what domain, and accordingly, they can use as per their requirements. Additionally, we have given a comparative view of the algorithms, which highlights the suitability of each algorithm in the respective domains. It also presents the advantages and disadvantages of each algorithm in several domains.

The rest of the sections are set out as follows: In Section 2, the DT induction algorithm is discussed and the classification tree techniques are explained in detail. Section 3 highlights the application details of the techniques explained in Section 2. A comparative analysis of various classification tree algorithms is presented in Section 4. In Section 5, the DT algorithms used for regression are explained in a simplified manner. Sections 6 and 7 incorporate application details and comparative analysis of the techniques reviewed under DT for regression, respectively. Sections 3-7 will help the beginners in deciding which algorithms to choose for their experimental works as they will get a broad perspective of the different techniques. Finally, in Section 8, the paper is concluded along with future works.

# 2 DT as a classifier

Classification is a way of fitting objects to a category which best suits its characteristics. Classification is a two-step process in which the first one constructs the classifier by examining vividly the training set containing the attributes and their associated class labels [25]. This step is called the training or learning phase [26] [27]. The second step is known as the classification phase where the performance of the classifier is measured for the testing dataset. If performance is found up to the mark, then those rules are applied to unknown data tuples to predict their class labels [28]. Classification intends to distinguish the discrete category of a new sample by contemplating a training dataset. Mathematically, the classification process can be presented as a function as follows [29]:

$$C = f(X, \theta), C \in L \tag{1}$$

where $X$ is the feature vector, $C$ is the class label of the new sample, $f(.)$ is the classification function, $\theta$ is the parameter set of the classification function and $L$, the set of class labels. The main objective of DT is to represent maximum possible training datasets correctly with the better performance [30]. The decision tree is constructed by observing the behavior of the training tuples. This procedure is known as decision tree induction [31]. The attribute values for a tuple whose corresponding class label is unknown are tested against the decision tree. In that way, the path traced from the root node to the leaf is used to obtain several possible intelligent classification rules.

The entire DT induction procedure is explained in Algorithm 1. The algorithm starts with a training set and an empty tree. In step 1, a single node N is generated. If instances are of the same class, then a node is appended to the tree containing that class (step 2). Step 3 illustrates the terminating condition. It says when the attribute list becomes empty, the leaf node of the DT contains the class label whose occurrence is highest. This is called the majority-voting approach [32]. Otherwise, the attribute that splits the dataset into best partitions is perceived using attribute selection methods (step 4). Steps 5 to 22 focus on the splitting criterion and possible subsets as a result of partitioning tuples as per the splitting criterion. While inducing a decision tree, the splitting criterion is the most important factor to be considered [33]. The splitting criterion helps us in choosing the attribute that divides the tuples in the dataset into partitions containing individual classes by making a test at node $N$. Hence, the split-point or the splitting subsets are determined according to the decision tree induction algorithm [34].

The dataset is partitioned with the aim that each of the partitions should be as pure as possible. If all samples in a partition of the dataset are linked to the same class, the partition is said to be pure. For an attribute $A$, having $x$ number of values $a_1, a_2, ..., a_x$, if it is discrete-valued, a set of branches are created corresponding to each attribute value. If it is continuous, then possible splits are in the

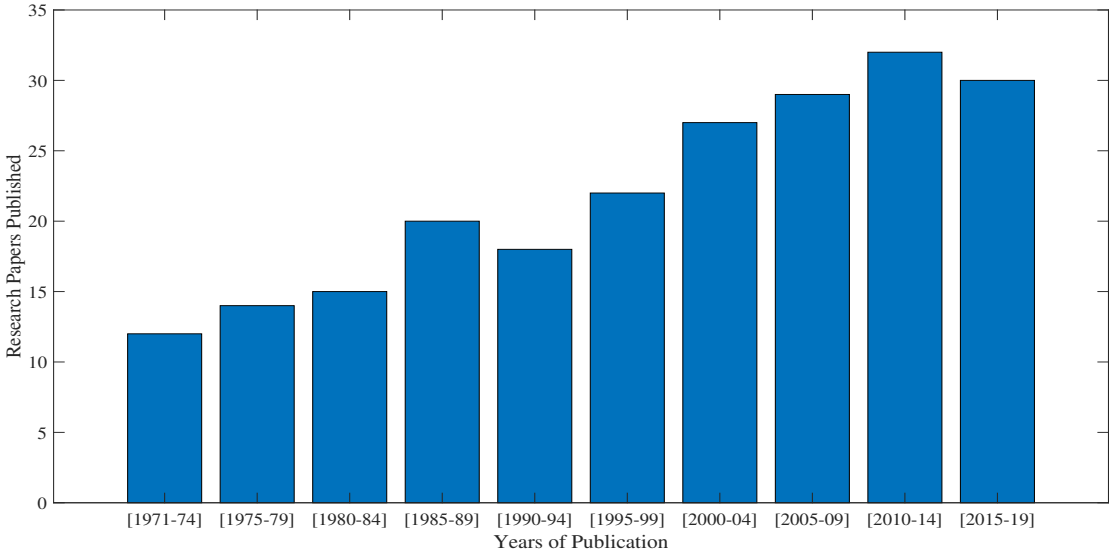



Figure 2: No. of research papers published over the years in the field of DT for Classification & Regression [Paper Sources: SCI, DBLP, Scopus indexed journals and conferences]

form of $a \leq c$ for one partition and $a > c$ for the other, where c is the splitting point. If the attribute is discrete and binary trees are to be generated only, then the splitting is in the form of $a \in S_a$, where $S_a$ is the splitting subset for attribute $A$. The scenario is depicted in Figure 3. Several decision tree algorithms have been proposed for the classification task of data mining by many pioneers in the field of machine learning and data mining. In this paper, we have discussed some of the popularly used algorithms and their working patterns.

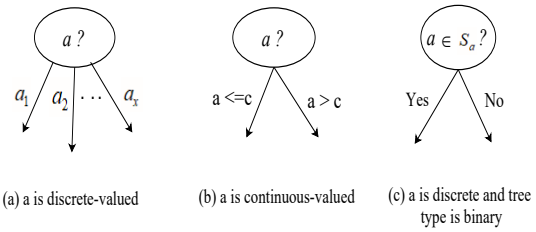



Figure 3: Different ways of partitioning tuples based on the splitting criterion

## 2.1 THeta automatic interaction detection (THAID)

This is the first published classification tree algorithm proposed by Messenger and Mandell [35]. It follows the concept of Automatic Interaction Detection(AID). AID is discussed in detail in section 5.1. THAID uses datasets having categorical variables. The node impurity at each node is measured based on the statistical distribution of the dependent variables over the mean. THAID searches the overall attributes of $X$ extensively and finds a set $S$, which reduces the node impurity of its children, then splits a node for the split $\{X \in S\}$. If $X$ is ordered, then $S \in (-\infty, c]$. Otherwise, $S \subseteq D(X)$, where $D(X)$ represents a set of possible values of $X$ (the domain of $X$). This procedure is repeated for the tuples in each child node and splitting halts when the relative decrease in node impurity becomes less than a predetermined threshold. THAID merges similar categories of the predictors for tree pruning.

## 2.2 CHi-squared automatic interaction detection (CHAID)

This algorithm is the extension of the AID approach, where the chi-square statistical test has been employed for finding the best split for each independent variable [36]. It was initially developed for classification and later extended to the task of regression. This algorithm can be applied to the samples having categorical, ordered with missing values, and ordered without missing values. CHAID performs better for categorical values in comparison to mixed mode data values. If the variables are continuous, they are converted to categorical before applying the CHAID algorithm. If the sample consists of ordered variables with $n$ distinct values, the chi-square test can be used to select the best suitable split out of $n - 1$ possible splits. If it consists of categorical variables and each variable is having $n$ categories, it can have $n$ splits. However, the number of splits can be lessened by applying Bonferroni adjusted significance tests. The significance test for each predictor follows a sequential cross-tabulation approach, whose steps are put forwarded in Algorithm 2. The major advantage of CHAID

**Algorithm 1** Decision Tree Induction Method

**Input:** Dataset S with attribute vector X $\{x_1, x_2, ..., x_n\}$ and each tuple in T$\{t_1, t_2, ..., t_p\}$ has associated class labels L$\{l_1, l_2, ..., l_m\}$
**Output:** Decision Tree
**Procedure:** DT_Induction

Generate a node N
**if** every $t_i$ in S $\in$ C **then**
  **return** N labeled with C
**end if**
**if** X = $\phi$ **then**
  **return** N as leaf with L=$max\{count(L_i)\}$ in S, $1 \leq i \leq m$
**else**
  find the best splitting criterion by applying attribute selection methods
**end if**
Label node N with attribute 'a' (the splitting attribute obtained from step 4).
**if** a is discrete and non-binary **then**
  X=X-a
**end if**
**for** each distinct outcome i $\in$ a **do**
  divide the dataset into $S_i$ partitions
  **if** $S_i = \phi$ **then**
    connect the leaf having $max\{count(L_i)\}$ to N, $1 \leq i \leq m$.
  **else**
    link the node returned by DT_Induction($S_i$, X) to N.
  **end if**
**end for**
**if** a is discrete and binary **then**
  X=X-$S_i$
  **for** each a $\in S_a$ **do**
    split at node N in such a manner that one split contains the tuples satisfying the condition and the other contains the remaining tuples.
  **end for**
**end if**
**if** a is continous **then**
  two splits are formed at split-point c
  Split A =$\sum_{i=1}^{p} t_i$, if $a > c$
  Split B =$\sum_{i=1}^{p} t_i$, if $a <= c$
**end if**
**if** partition is not pure or spilitting is further Possible **then**
  **goto** ***Start***
**end if**
return N

is, it reduces the computational complexity by reducing the number of categories for each predictor using the merging procedure.

**Algorithm 2** Sequential Cross-Tabulation Approach

1: Cross-tabulate $n$ categories of independent variables with m categories of dependent variables.
2: Apply the chi-square test on the cross table and find the pair of categories of the independent variables which are least significantly different.
3: Merge the two categories which pass through step 2.
4: Repeat steps 2 and 3 until no non-significant chi-square test result is obtained.
5: Select the attribute whose chi-square result is largest, and split into $k$ branches where, $k \leq l$, and $l$ is the number of categories of the independent attributes obtained from the merging process.
6: Repeat step 5 until the stopping criteria is satisfied.

## 2.3 Iterative dichotomizer(ID3)

It employs entropy as a measure of node impurity [37] [38]. It uses ordered discrete attributes. The expected information or entropy relies on the probability of belongingness ($P_i$) of any tuple of a dataset $D$ to a particular class. Entropy for $n$ classes in a dataset can be computed as follows [17]:

$$En(D) = -\sum_{i=1}^{n} P_i \log_2(P_i), \tag{2}$$

where, $P_i = \frac{|S_i|}{|S|}$ , in which the denominator denotes the number of tuples in $D$ and the numerator contains the amount of samples with respect to class $C_i$. In addition, the entropy of the partitions is to be calculated based on the values of attribute $t$ in the dataset $(D)$. For $s$ distinct values $\{t_1, t_2, t_3, ..., t_s\}$ of each attribute $t$, the entropy of the partition with respect to $t$ is:

$$En_t(D) = \sum_{i=1}^{s} \frac{|D_i|}{|D|} \times En(D_i). \tag{3}$$

where $D$ is partitioned into $s$ subsets $\{D_1, D_2, D_3, ..., D_s\}$, and $En(D_i)$ is the entropy of the partition with respect to values of an attribute *t*. $D_i$ consists of the tuples in $D$ having outcome $t_i$ of the attribute *t*. This is required to obtain the exact classification of the instances. The information gain, $G(t)$, is computed as follows:

$$G(t) = En(D) - En_t(D). \tag{4}$$

The attribute with highest $G(t)$ or minimum $En_t(D)$ is chosen as the splitting attribute. Originally, ID3 was proposed considering discrete data only, but later it experimented on continuous data in several works. Some have

considered the midpoint between each pair of adjacent values as a possible split-point and some have used discretization to convert continuous data to discrete and then applied ID3 on that data. In case of midpoint procedure, possible splits for an attribute $t$, are of the form $t \leq c$ for one set of tuples, and $t > c$ for another set of tuples where, $c$ is the split-point between two adjacent pair of attribute values $t_i$ and $t_{i+1}$. The value of $c$ can be calculated as: $(t_i + t_{i+1})/2$.

## 2.4 C4.5

C4.5 is a descendant of ID3, proposed by J. R. Quinlan [39]. The major limitation of ID3 is that it gives preference to the attributes having more values and more missing values. In order to overcome this problem, gain ratio was adopted as the attribute selection measure instead of entropy. For $s$ subsets $D_1, D_2, ...., D_s$ of dataset $D$, instead of using the entropy, it uses the splitting information $(SI_t)$ [17]:

$$SI_t(D) = \sum_{i=1}^{s} \frac{|D_i|}{|D|} \times log_2 \left( \frac{|D_i|}{|D|} \right). \tag{5}$$

The gain ratio ($GR$) is the ratio of entropy and $SI_t(D)$:

$$GR(t) = \frac{G(t)}{SI_t(D)} \tag{6}$$

The attribute having the highest $GR(t)$ value is chosen as the splitting attribute. The problem arises when $SI_t(D)$ becomes negligible or tends to zero. It leads to unbalanced ratio; hence one constraint needs to be imposed, that is, $G(t)$ value should be large enough when the gain ratio is applied.

## 2.5 Classification and regression trees (CART)

In contrast to ID3 and C4.5, it generates binary decision trees [40]. It works on both discrete and continuous data. It uses gini index (GI) as a measure of node impurity [41].

$$GI(D) = 1 - \sum_{i=1}^{n} P_i^2, \tag{7}$$

where, $P_i = \frac{|S_i|}{|S|}$ is the ratio of number of tuples present in the dataset with respect to a particular class to the total number of tuples present in D. For a binary split with respect to an attribute 't', GI can be calculated as:

$$GI_t(D) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} GI(D_i) \tag{8}$$

where, $D_i$ is the gini index with respect to a partition. Due to the binary split on attribute t, the reduction in impurity is computed as:

$$GI_{red}(t) = GI(D) - GI_t(D) \tag{9}$$

For each attribute, every feasible binary splits are taken into consideration. The subset with minimum $GI_{red}(t)$ is chosen as the splitting subset [42]. For continuous-valued attributes, it uses the same midpoint procedure as ID3 to find a possible split-point.

## 2.6 Fast and accurate classification trees (FACT)

The FACT algorithm for decision tree used for classification is similar to the recursive Linear Discriminant Analysis (LDA) procedure in which the tree is constructed with linear splits [43]. The number of children for each predictor is the same as the number of classes for that variable. In this algorithm, the predictors are being ranked based on the Analysis of Variance (ANOVA) and F-test, and the splitting procedure is performed on the selected predictor based on the LDA method [44]. Initially, all the categorical independent variables are transformed into ordered variables using an intermediate binary vector. One of the specialties of this algorithm is the procedure of handling the missing values. It estimates the means and modes of non-missing data values of ordered and categorical predictors respectively, and replaces those values in place of missing values. The size of the tree is identified based on the stopping criteria of the ANOVA test [45]. The major advantage of this algorithm is, it is unbiased towards the selection of predictors at each level. However, it is biased towards the predictors, which are categorical as LDA is employed to convert it into an ordered one. This limitation is addressed by the Quick Unbiased Efficient Statistical Tree (QUEST) algorithm, which removes the bias for the splitting of ordered variables.

## 2.7 Quick unbiased efficient statistical tree (QUEST)

It is an efficient decision tree classifier that addresses the FACT algorithm's limitation, which is biased towards the selection of categorical variables. QUEST uses the cross-tabulation approach of chi-squared tests, and F-tests to handle categorical and ordered predictors, respectively to give a fair chance of selection [46]. When a binary split is required at a node with more than one class, it merges the classes into two superclasses before the significance test is applied. If the variable is ordered, the split-point is chosen by quadratic discriminant analysis or the exhaustive search. Apart from that, if the variable is categorical, the point of splitting is chosen after transforming it into a larger discriminant coordinate. The major advantage of QUEST is, it improves the computational time over CART when variables with many categories exist.

### 2.8 Classification rules with unbiased interaction selection and estimation (CRUISE)

In CRUISE, each node is split into multiple branches, which depends on the number of class labels associated with the independent variables [47]. It is an extension of QUEST. The variable selection at each level is based on the cross-tabulation approach used in CHAID, where the columns and rows of the cross table contain the predictors and class labels, respectively. Unlike QUEST, it performs the significance tests between two independent variables, say $X_i$ and $X_j$ instead of performing pairwise significance tests between two categories of $X$ variables [48]. If the significance test between $X_i$ and $X_j$ are found to be best, $X_i$ is chosen for splitting instead of $X_j$. The split-point is then identified by the LDA approach after the independent variable goes through a Box-Cox transformation. The major advantage of CRUISE is, it allows splitting of all the variables linearly that can fit the LDA model at each leaf node. Another advantage is, it is unbiased towards the selection of variables that have more missing values.

### 2.9 Generalized unbiased interaction detection and estimation (GUIDE)

GUIDE is the improvised version of QUEST and CRUISE. It models the decision tree classifier by leveraging the strengths of both algorithms. It also reduces the limitations of CRUISE by minimizing the number of interaction tests among the categorical variables. The amount of computation is drastically reduced as it restricts the frequency of tests. The multi-level searching technique is employed for splitting at each node when the significant difference between two variables $X_i$ and $X_j$ is noticed. The first level splitting of a node is performed based on $X_i$ and the second level splitting based on $X_j$ in order to reduce the amount of impurity. This process is repeated in a reverse manner, i.e, $X_j$ is considered for splitting at first level and $X_i$ in second. The one whose reduction in impurity is greater is chosen to split the node. One of the advantages of GUIDE is, it can perform bivariate splits of two independent variables at a time along with univariate splits. Bivariate linear split is preferable over univariate if the number of observations at each node is found to be lesser than the number of independent variables.

### 2.10 Conditional inference tree (CTREE)

It can handle ordered, nominal, continuous, censored as well as multivariate attributes. It uses the combination of recursive binary partitioning and theory of permutation to select split variables [49]. Based on Bonferroni adjusted p-values, it derives stopping rules to regulate the tree size instead of applying tree pruning to reduce the tree size. Like CART, it also uses surrogate splits to deal with missing values, and the number of surrogate splits can be regulated by defining maximum surrogate splits using a function.

## 3 Application details of the techniques reviewed under DT for classification

This section exemplifies a brief illustration of the splitting criterion used, application areas, dataset details, and performances of the different algorithms reviewed under DT for classification. THAID was used in finance and health care for various purposes. CHAID was applied in many application areas like marketing, health care, coal mining, etc and its performance is comparable with several algorithms of its time. CHAID was also used in the public vocational rehabilitation program to predict the employment outcomes and acceptance rates of rehabilitation clients with orthopedic disabilities. ID3 is adopted in many application areas like price prediction in stock markets, in health care for medical diagnosis and it is having better classification accuracy than neural networks and rough sets classifiers. The extended version of ID3, i.e., C4.5 was employed in several sectors like health care for liver disease diagnosis, detection of cancer disease with the help micro-array datasets, and tumor classification [50]. It is also used in land cover mapping and change assessment in remote sensing, etc. Its performance is comparable with k-Nearest Neighbor (kNN), Naive-Bayes, and Support Vector Machine (SVM) classifiers.

Similarly, CART is used in various fields like intrusion detection, bankruptcy prediction in companies [51]; diagnosis of diabetes and prediction of heart disease in health care [52]; landslide hazard, etc and have shown better performance than ID3. Likewise, FACT is also used in many areas like waveform recognition, digit recognition, and normal discrimination. Researchers employed QUEST in educational institutions for evaluating teachers' performance [33], in health care for predicting mortality rate because of head injury, financial firms for measuring firm performance, etc. Likewise, GUIDE, CRUISE, and CTREE are used in several research areas and are efficient and effective for the researchers. The details are mentioned in Table 1.

## 4 Comparative analysis of various classification tree algorithms

In this section, different classification tree techniques, as discussed, are compared based on various parameters, as listed in Table 2. The parameters considered for the comparison are different types of splits (univariate or linear), the maximum number of splits, the way they handle missing valued attributes, node models, etc. CHAID and C4.5 algorithms do not support linear splits. However, most of the algorithms support both linear and univariate splits. The prediction accuracy of THAID is not up to the mark.

Table 1: Application details of the techniques under DT for Classification

| Sl No. | Method | Splitting Criterion | Application Area | Dataset Details | Remarks |
|---|---|---|---|---|---|
| 1 | THAID | Sum of Squared Deviation | Finance, Health care | Car dataset from 1970 survey of Consumer finances, IRIS dataset from UCI repository, | Low predictive accuracy, Biaseness in variable selection |
| 2 | CHAID | Chi-squared Statistical test | Marketing modelling. Healthcare, Coal mining | IRIS dataset, Breast cancer patients' data, Coal mines data from Coal Industry Promotion Board, Rehabilitation Service Administration (RSA)-911 dataset | Performance comparable and in many cases outperforms other algorithms, restricted to categorical variables |
| 3 | ID 3 | Entropy | Product entry decision, Weather forecasting, Medical diagnosis, Marketing, Stock market trend mining | Heart disease data from UCI rep., Weather data, Buys_computer data | Predictive accuracy is directly proportional to the size of the training set; Better classification accuracy than rough sets and neural networks |
| 4 | C 4.5 | Gain Ratio | Finance, Health care, Land cover change assessment | Car dataset from Journal of Statistics Education Data Archive, IRIS dataset from UCI repository, Liver Disorders datasets from UCI repository , data sets Landsat 5 (TM) for 1986 and Landsat 7 (ETM+) for 2001 located on the satellite path; Leukemia, Colon tumour and Diffuse Large B-cell Lymphoma data from Kent Ridge Bio-Medical Data Set Repository | Performance comprable with classifiers SVM, k-NN, Naive Bayes' |
| 5 | CART | Gini Index | Medicine and Health care, Landslide hazard, Intrusion Detection | Car dataset, Birth dataset, Type 2 Diabetic outpatient data, Survey data of malaria in central vietnam during 2008, KDD Cup 1999 dataset from UCI rep., Landslide data set of 137570 samples from Penang Island in Malaysia | Great flexibility and accuracy but splitting is biased towards variables having more distinct values |
| 6 | FACT | ANOVA and f-test | Normal discrimination, Digit recognition, Waveform Recognition, Spherical distribution problem | IRIS dataset, Boston housing dataset | Classification accuracy and interpretative capability is comparable with CART, but FACT runs many times faster |
| 7 | QUEST | Chi-squared & f-test | Financial firms, Health care, Landslide hazard, Coal mine | Car dataset from Journal of Statistics Education Data Archive, IRIS dataset, Financial data of Turkish firms from FINNET, Breast cancer patients' data, Coal mines data from Coal Industry Promotion Board | Unbiased splits, ranked fourth best overall for linear splits, Improved computational time over CART for variables of many categories |
| 8 | CRUISE | LDA, Contingency Table Chi-squared tests | Biomedicine, Education, Healthcare | IRIS dataset, Biomedical data, Cylinder bands, Credit approval, Echo-cardiogram, Fish catch, Horse colic, Hepatitis, Heart disease, Auto imports from UCI rep.; Demography data from Rouncefield (1995), Head injury from Hawkins(1997), College data from StatLib | Accuracy as high as CART and QUEST, fast computation speed, produces more intelligent splits and shorter trees, keeps track of local interactions |
| 9 | GUIDE | Bonferroni test, Chi-squared test | Education, Sports, Healthcare Region prediction | IRIS dataset from UCI, Cars dataset from the Journal of Statistics Education Data Archive for 2004 model year | Performance better than CRUISE and QUEST, Unbiased variable selection |
| 10 | C-TREE | Bonferroni p-test | Healthcare, Sports, Space Physics, Mammography, Biology | Breast cancer, Credit, Heart, Hepatitis, Ionosphere, Sonar, Liver, TicTacToe, Titanic House votes 84 from UCI repository | Performance comparable and in some cases better than GUIDE, Unbiased, uses permutation tests |

CHAID favors categorical variables, and it allows multiple splits at a node. CART has great flexibility and accuracy, but splitting is biased towards variables having more distinct values. The classification accuracy and interpretative capability of FACT are comparable with CART, and it runs many times faster than CART. CART, CHAID, and QUEST are the most popular techniques used for modeling decision trees for classification. The QUEST algorithm is a little bit faster as compared to CART and CHAID. However, it is not suitable for processing bigger datasets as it requires high storage space to store the intermediate results obtained at each level of the tree.

QUEST, CRUISE, GUIDE, and CTREE are the advanced approaches to model the classification tree. They were found effective in terms of both time and space complexity. They also provide unbiased splits during the construction of the classification tree. Accuracy of CRUISE is as good as CART and QUEST; it has fast computational speed, generates shorter trees, more intelligent splits, and also keeps track of local interactions that makes it distinguishable from other algorithms proposed before it [18]. GUIDE is having better accuracy than CRUISE and

Table 2: Comparision of classification tree algorithms

| Author Name | Year | Algorithm | Split type | Unbiased Split | No. of splits | Missing values Method | Interaction Test | Node Model |
|---|---|---|---|---|---|---|---|---|
| R. Messenger & L. Mandell | 1972 | THAID | U | No | 2 | – | Yes | C |
| G. V. Kass | 1980 | CHAID | U | No | >=2 | B | Yes | C |
| J R Quinlan | 1986 | ID3 | U | No | >=2 | – | No | C |
| J R Quinlan | 1993 | C 4.5 | U | No | >=2 | W | No | C |
| L Breiman et al. | 1984 | CART | U,L | No | 2 | S | No | C |
| W Y Loh & N. Vanichsetakul | 1988 | FACT | U,L | No | >=2 | I | No | C |
| W Y Loh & Y S Shin | 1997 | QUEST | U,L | Yes | 2 | I | No | C |
| H Kim & W Y Loh | 2001 | CRUISE | U, L | Yes | >=2 | I, S | Yes | C, D |
| W Y Loh | 2002 | GUIDE | U, L | Yes | 2 | M | Yes | C, K, N |
| T Hothorn et al. | 2006 | C-TREE | U,L | Yes | >=2 | I,S | No | C |

*Description: U- univariate splits, L- Linear splits, B-Missing value branch, W- Probability weights, S- Surrogate splits, I- Missing value imputation, C- Constant model, M- missing value category, D- Discriminant model, K- kernel density model, N- Nearest neighbour model. Blank entries indicate 'no missing values'* .

QUEST and is having an unbiased variable selection. In contrast to others, CTREE uses permutation tests. Its performance is comparable, and in some cases, it is better than GUIDE.

# 5 DT for regression

Regression aims to predict a continuous value for an unseen tuple by studying a training sample of data [29]:

$$O = f(x, \theta), O \in R \tag{10}$$

where, $x$ is the new observation, $O$ is the output, $f(.)$ is the regression function and $\theta$ is the regression function's parameter set. DT for regression is similar to classification trees with the difference that it contains values or piecewise models at leaves rather than class labels [53]. The values may be the result of any test or the outcome of any operation. Some of the popularly used regression tree algorithms are discussed in this section.

## 5.1 Automatic interaction detection (AID)

It is the first regression tree algorithm, introduced by Morgan and Sonquist in the year 1963. This algorithm starts with a large dataset. The large dataset is then successively divided into several subgroups after applying binary divisions. At every step, the binary divisions of the groups are defined by one of the independent variables. It uses the sum of squared deviations as a measure of node impurity [54]. For each independent variable, all possible splits are considered. Each binary split divides the whole dataset into two parts. The one having least sum of squared deviations is chosen. The node impurity measure ($I(d)$) is computed as follows [24]:

$$I(d) = \sum_{i=1}^{n} (y_i - \bar{y_d})^2 \tag{11}$$

where, $\bar{y_d}$ is the sample mean of dependent variables with respect to the partition. The attribute with the least sum of squared deviations is taken as the splitting attribute. The splitting process continues till very few tuples remain in the dataset or when $I(d)$ becomes less than a predefined value. The task of deciding the predefined value is a matter of concern, as it might lead to the problem of over-fitting or under-fitting if the number is either too large or too small, respectively. Inter-correlation among attributes leads to spurious results. A biased value is considered during the model building process.

## 5.2 CART for regression

It uses the same approach as AID for splitting and computing the node impurity measure. It solves the over-fitting problem of AID by using the tree pruning procedure. The yield of CART is piecewise constant models. CART uses *surrogate splitting* approach to handle datasets with missing values [55]. If splitting needs to be performed on an attribute with missing values, then it finds an attribute that is highly correlated to the original attribute and replaces that attribute with the original one.

## 5.3 Multivariate adaptive regression splines (MARS)

MARS is suitable for handling datasets of higher dimensions. It follows the recursive, divide and conquer approach as regression and generates continuous models with continuous derivatives [56]. It splits the range of independent

attribute values into $n+1$ disjoint intervals partitioned by $n$ knots, which results in the construction of functions, called spline functions [57]. MARS comprises of a series of connected straight line segments. The general form of MARS model is defined as [58]:

$$y = f(x) = z_0 + \sum_{i=1}^{n} z_i B_{kn}(x_{v(k,i)}) \tag{12}$$

where, $y$ is the output function, $n$ is the number of basis functions, $z_0$ is a constant value, $k$ is the order of interactions, $x_{v(k,i)}$ is the independent attribute in the $k^{th}$ of the $i^{th}$ product, $B_{k,n}(x_{v(k,i)})$ is the $i^{th}$ basis function and $z_i$ is its corresponding coefficient. The basis function can be defined as: $B_{kn} = \prod_{i=1}^{k} b_{in}$. The value of $k$ is one if the model is additive, and it is two, for the pairwise interactive model. In the first step, a significant quantity of basis functions are constructed which overfit the data. The permitted data values are categorical, continuous, and/or ordinal and they are selected as per the intervals defined. The different variables may have direct interaction with each other or some constraints may be imposed on them. In the second phase, a generalized cross validation technique is applied on the basis functions and the functions having the least contribution are eliminated. The variables having better cross-validation results are chosen. In this way, an optimal MARS model is selected. MARS successfully handles missing values by employing dummy variables. By using the above-mentioned procedures, MARS also keeps track of complex data structures hidden in high dimensional datasets.

## 5.4 GUIDE for regression

It employs the chi-square test to detect the interrelationships between the signed residuals and groups of independent variables [59]. It can handle datasets having both discrete and continuous-valued attributes. In GUIDE, two tests are performed, curvature test and interaction test. In the curvature test, for each continuous-valued attribute, a 2 x 4 table, called contingency table is created using the dataset, whose rows indicate signs of the residuals and columns stipulate groups. Based on the number of observations in each cell, the p-value is obtained from the chi-square distribution. In the Interaction test, to find interaction among two continuous variables, the sample median is computed, and based on the result, the range of each variable is divided into two equal partitions. A $2 \times 4$ contingency table is generated, whose rows represent residual signs and columns denote quadrants. The chi-square distribution and p-value are also computed in this algorithm. If the acquired p-value is a consequence of the curvature test, the corresponding independent variable is chosen as the splitting attribute; and if it is from the interaction test, then one of the interacting variables is chosen as the splitting attribute. The sum of squared error is computed for each sub-node, and the variable having the least sum of squared error is selected. In case one of the variables is categorical, the one having a smaller p-value as a result of the curvature test is selected. The major advantage of GUIDE is that it is unbiased towards the splitting process.

## 5.5 M5

It involves the construction of model trees rather than the rule based, recursive binary trees [60]. Model trees are smaller in structure than regression trees and have shown better performance than the later. They can handle datasets of large dimensions. In contrast to regression trees, which contain values at their terminal nodes, model trees employ linear functions. M5 can handle both discrete and continuous data. Its objective is to build a model that associates the target values of the dependent variables to other attributes' values [61]. The construction of model trees follows the divide and conquer approach. If the constructed model suffers from over-fitting, tree pruning is applied by substituting a subset with a leaf. M5 considers standard deviation as a measure of node impurity. At first, the standard deviation of the dependent variables is computed in the training sample of the dataset ($Dt$). Based on the outcomes of the test, the splitting process continues until there is no notable distinction between the values of the attributes. By ascertaining the subset of data tuples associated with each outcome, every potential test is evaluated. The expected reduction in error can be calculated as [60]:

$$\Delta err = \sigma(Dt) - \sum_{i=1}^{n} \frac{Dt_i}{Dt} \times \sigma(Dt_i) \tag{13}$$

where, $Dt_i$ denotes the subset of data tuples having $i^{th}$ outcome of the potential test, n denotes the number of outcomes of a test, and $\sigma(Dt)$ represents standard deviation of the training dataset. The test having maximum $\Delta err$ is chosen as the potential test to predict the target values of the unseen data tuples. The test set error ($\Delta err_t$) can be computed as:

$$\Delta err_t = \Delta err \times (\frac{m+p}{m-p}) \tag{14}$$

where $m$ represents the number of training set tuples at a particular node, and $p$ refers to the number of parameters in the regression model of the node.

## 5.6 $M5'$

It is an extension of M5, designed to address some issues that arose during the construction of M5. It is a k+1 parameter model, where k attributes and one constant term $w_0$ are there. In M5, as the size of the tree becomes smaller, the standard deviation in $\Delta err$ lessens. Hence, to manage this, a pruning factor called $\alpha$ is used in $M5'$ while computing $\Delta err_t$ [62].

$$\Delta err_t = \Delta err \times (\frac{m+\alpha p}{m-p}) \tag{15}$$

As the $\alpha$ value increases, $\Delta err_t$ increases but the size of the tree decreases outstandingly. Hence, to get less error and better performance, a smaller value of $\alpha$ must be taken; but if preference will be given to generate smaller trees then $\alpha$ value must be increased a little bit. In addition to this, $M5'$ also successfully handles the datasets containing missing values [63]. To address missing values, some modifications to $\Delta err$ has been done as follows [62]:

$$\Delta err = \frac{k}{|Dt|} \times \beta(i) \times \left[ \sigma(Dt) - \sum_{j \in \{A,B\}} \frac{|Dt_j|}{|Dt|} \times \sigma(Dt) \right] \tag{16}$$

where k refers to the number of tuples without missing values, Dt is the subset of the dataset containing tuples that are to be split based on a condition, $Dt_A$ and $Dt_B$ are the sets after partition, and $\beta(i)$ is the correction factor defined as [62]: $\beta = e^{7 \times \frac{2-a}{m}}$, where m is the number of tuples in the dataset, and a is the total number of values of original enumerated attributes. $\beta$ is used for converting 'a' valued enumerated attributes to 'a-1' binary values. For continuous attributes, $\beta$ is taken as 1. Hence, after splitting, all attributes in $Dt_A$ and $Dt_B$ become binary. The attribute with a maximum $\Delta err$ is chosen as the splitting attribute. The prediction accuracy of $M5'$ is comparable with techniques like Artificial Neural Networks(ANN) and is found to be better than that of regression trees like CART [64].

# 6 Application details of the techniques reviewed under DT for regression

This segment epitomizes a short depiction of the different algorithms reviewed under DT for regression based on few parameters as alluded in Table 3. It starts with AID. It has been adopted in several research areas like education for predicting the factors affecting the academic survival of students, in population studies for the adoption of family planning in Koyang [65], in marketing for exploratory analysis of market data, etc. CART is used in several fields like sports for analyzing the salary of baseball players [24], in public health for analyzing causes of morbidity and mortality from specific diseases, and in many areas for different tasks using several datasets [53]. MARS is adopted in several research areas like health care on heart attack survival data, in biology for prediction of species distributions, etc. In some applications like credit scoring, CART and MARS outperform traditional logistic regression, discriminant analysis, SVM, and neural network techniques in terms of predictive accuracy. GUIDE is employed in various sectors like education, sports, and automobiles. MARS and M5 were also used for groundwater level forecasting, solar radiation, in the construction industry for evaluating mechanical properties of concretes containing coarse recycled concrete aggregates [66] and they have outperformed other algorithms in many scenarios. $M5'$ is used in various application areas like coastal engineering for prediction of wave height in Lake Superior [67], for scour depth prediction [68], in construction industry for predicting modulus elasticity of recycled concrete [63], etc. For more details, Table 3 may be referred.

# 7 Comparative analysis of regression tree algorithms

Comparative analysis of various regression tree algorithms based on different parameters is presented in Table 4. Some of the parameters considered for comparison are the same as the parameters used for comparing classification tree algorithms. However, few parameters like pruning, variable importance ranking, loss criteria, ensemble approach are added for an extensive comparison of regression tree algorithms. Regardless of its novelty, AID faced some problems and was criticized by several authors [24]. While splitting, it experiences over-fitting as well as under-fitting. It doesn't employ tree pruning to reduce the tree size, whereas CART and others do the same to reduce the complexity. The square of mean deviation is considered as the node impurity adopted in the AID and CART algorithm. MARS employs a spline basis function and incorporates a generalized cross-validation approach that increases the prediction accuracy of the model. GUIDE employs ensemble and bagging techniques in contrast to others. $M5'$ is the best regression model which constructs the piecewise constant tree by fitting the linear regression model at each leaf node whereas, the GUIDE algorithm fits linear regression models at each node in the constructed tree. For detailed analysis, Table 4 may be referred.

# 8 Conclusions and future work

The popularity of the classification and regression trees has been increasing exponentially, as they are easy to understand and implement. In the decision tree, the hidden rules along with the constraints can be extracted from the data and can be mapped with the nodes and branches of the tree, which makes it more convenient for understanding. However, the complexity of the model increases with the increase in the size of the datasets. To handle the complexity, a wide number of advanced algorithms have been adopted in the field of DT for classification and regression. In this paper, we have presented the list of the datasets and various applications in which these algorithms can be applied. This paper could be a potential resource for the researchers in searching and deciding the appropriate algorithms suitable for their area of research, which involve regression and classification task. The comparative analysis of numerous algorithms based on various parameters is also presented for both classification and regression tasks. In future, this work can be extended by including all the en-

Table 3: Application details of the techniques under DT for Regression

| Sl. No. | Method | Splitting Criterion | Application Area | Dataset Details | Remarks |
|---|---|---|---|---|---|
| 1 | AID | Sum of Squared Deviations | Education, Population Studies, Market Research, Operational Research, Fishing Industry, Gasoline Consumption | Data from almost 6,000 gas stations from major oil companies in the Unites States during 1970; The shing log data from The White Fish Authority, Hull, England; Data taken from Ronald Freedman, P. Whelpton and Arthur Campbell; Family Planning, Sterility and Population Growth(New York, 1959); Data from NestleCompany, Edu- cational data from The National Survey of Health and Development | It is biased towards datasets of higher dimensions, Experiences overfitting and underfitting problems |
| 2 | CART | Gini Index | Medicine and Health Care, Landslide hazard | Baseball salary data from American Statistical Association Section(StatLib), Data from the 1999 Behavioral Risk Factor Surveillance System (BRFSS) (61), conducted annually by U.S. states' Departments of Health in collaboration with the Centers for Disease Control and Prevention | Great flexibility and accuracy but splitting is biased towards variables having more distinct values |
| 3 | MARS | Spline Basis Functions, Generalized Cross Validation | Health Care, Biology, Credit Scoring, Solar Radiation, Construction Industry | Heart attack survival data from Specialized Center of Research on Ischemic Heart Disease at the University of California, San Diego; Birds data of many countries and Plants data of Switzerland; Credit card data set provided by a local bank in Taipei, Taiwan, Solar data from Data obtained from Adana and Antakya stations, Turkey | Incorporation of Generalized Cross Validation increases the prediction accuracy, Handles Curse of dimensionality problem |
| 4 | GUIDE | Chi-squared test of interaction | Sports, Automobiles, Education | Baseball salary data from StatLib; Car dataset from the Journal of Statistics Education Data Archive for 2004 model year | Fast computation speed, Unbiased and keeps track of local interactions during split selection |
| 5 | M5 | Standard deviation | Medicine and Health Care, Manufacturing, Automobiles, Hydrology, Solar Radiation, Evapotranspiration | CPU performance data; Car price data; Drug Activity data, LHRH data from Arris Pharmaceuticals, San Francisco; Data from a discharge measuring station Swarupganj on the river Bhagirathi, India; Sediment yield data from Nagwa watershed in India from 1993 to 2004; Solar data from Data obtained from Adana and Antakya stations, Turkey, Climatic data of Davis station maintained by California Irrigation Management Information System (CIMIS) | Better accuracy and smaller in structure than regression trees |
| 6 | M5' | Standard deviation | Marine & Coastal Engineering, Construction Industry, Coastal and Ocean engineering | Wind and Wave data gathered in Lake Superior from 6 April to 10 November 2000 and 19 April to 6 November 2001; wave run-up data of Van der Meer and Stam (1992) | The prediction accuracy is comparable with techniques like Artificial Neural Networks and is found to be higher than CART, Handles datasets with missing values |

Table 4: Comparison of regression tree algorithms

| Author Name | Year | Algorithm | Split Type | Unbiased Split | Number of Splits | Pruning | Variable Importance Ranking | Node Models | Missing value methods | Loss Criteria | Bagging and Ensembles |
|---|---|---|---|---|---|---|---|---|---|---|---|
| J. N. Morgan and J. A. Sonquist | 1963 | AID | U | No | 2 | No | Yes | C | – | V | No |
| L Breiman et al. | 1984 | CART | U,L | No | 2 | Yes | Yes | C | S | V | No |
| J. H. Friedman | 1991 | MARS | L | Yes | >=2 | Yes | Yes | C, M | A | V | No |
| W Y Loh | 2002 | GUIDE | U | Yes | 2 | Yes | Yes | C, M, P, R | A | V, W | Yes |
| J. R. Quinlan | 1992 | M5 | U | No | >=2 | Yes | No | C, R | – | V | No |
| Y. Wang and I. H. Witten | 1996 | $M5'$ | U | No | >=2 | Yes | No | C, R | G | V | No |

*Description: U-Univariate splits, L-Linear splits, C-Constant Model, M-Multiple linear model, R- Stepwise linear model, P- Polynomial Model, S- Surrogate splits, G- Global mean/mode imputation, A- missing value category, V- Least Square, W- Least Median square [Blank entries in the table indicate those algorithms do not handle datasets with missing values]*

semble approaches and their comparison with the existing ones. We also aim to explore new techniques in the field of decision tree-based hierarchical multi-label classification, multi-output, and multi-objective regression trees, etc.

# References

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding (2013) Data mining with big data, *IEEE transactions on knowledge and data engineering*, 26(1), pp. 97–107. https://doi.org/10.1109/tkde.2013.109

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami (1993) Database mining: A performance perspective. *IEEE transactions on knowledge and data engineering*, 5(6), pp. 914–925. https://doi.org/10.1109/69.250074

[3] Ranjan Kumar Behera, Santanu Kumar Rath, Sanjay Misra, Robertas Damaševičius, and Rytis Maskeliūnas (2017) Large scale community detection using a small world model. *Applied Sciences*, 7(11),

pp. 1173.
https://doi.org/10.3390/app7111173

[4] Satchidananda Dehuri and Ashish Ghosh (2013) Revisiting evolutionary algorithms in feature selection and nonfuzzy/fuzzy rule based classification, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(2), pp. 83–108. https://doi.org/10.1002/widm.1087

[5] Leszek Rutkowski (2004) Adaptive probabilistic neural networks for pattern classification in time-varying environment, *IEEE transactions on neural networks*, 15(4), pp. 811–827. https://doi.org/10.1109/tnn.2004.828757

[6] Ranjan Kumar Behera, Debadatta Naik, Dharavath Ramesh, and Santanu Kumar Rath (2020) Mr-ibc: Mapreduce-based incremental betweenness centrality in large-scale complex networks, *Social Network Analysis and Mining*, 10, pp. 1–13. https://doi.org/10.1007/s13278-020-00636-9

[7] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens (2011) Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert systems with applications*, 38(3), pp. 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

[8] Charu C Aggarwal (2014) *Data classification: algorithms and applications*, CRC press.

[9] Salvador García, Alberto Fernández, and Francisco Herrera (2009) Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems, *Applied Soft Computing*, 9(4), pp. 1304–1314. https://doi.org/10.1016/j.asoc.2009.04.004

[10] Shih-Wei Lin, Kuo-Ching Ying, Chou-Yuan Lee, and Zne-Jung Lee (2012) An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection, *Applied Soft Computing*, 12(10), pp. 3285–3290. https://doi.org/10.1016/j.asoc.2012.05.004

[11] Sreerama K Murthy (1998) Automatic construction of decision trees from data: A multi-disciplinary survey, *Data mining and knowledge discovery*, 2(4), pp. 345–389.

[12] Arno De Caigny, Kristof Coussement, and Koen W De Bock (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, *European Journal of Operational Research*, 269(2), pp. :760–772. https://doi.org/10.1016/j.ejor.2018.02.009

[13] Usama M Fayyad and Keki B Irani (1992) On the handling of continuous-valued attributes in decision tree generation, *Machine learning*, 8(1), pp. 87–102. https://doi.org/10.1007/bf00994007

[14] Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski (2007) Ensembles of multi-objective decision trees, *European conference on machine learning*, Springer, pp. 624–631. https://doi.org/10.1007/978-3-540-74958-5_61.

[15] Dua Dheeru and Efi Karra Taniskidou (2017) UCI machine learning repository.

[16] Jieyue He, Hae-Jin Hu, Robert Harrison, Phang C Tai, and Yi Pan (2006) Transmembrane segments prediction and understanding using support vector machine and decision tree, *Expert Systems with Applications*, 30(1), pp. 64–72. https://doi.org/10.1016/j.eswa.2005.09.045

[17] Jiawei Han, Jian Pei, and Micheline Kamber (2011) *Data mining: concepts and techniques*, Elsevier.

[18] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine learning*, 40(3), pp. 203–228.

[19] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman (2002) Decision trees: an overview and their use in medicine, *Journal of medical systems*, 26(5), pp. 445–463. https://doi.org/10.1023/a:1016409317640

[20] Claudia Perlich, Foster Provost, and Jeffrey S Simonoff (2003) Tree induction vs. logistic regression: A learning-curve analysis, *Journal of Machine Learning Research*, 4(Jun), pp. 211–255.

[21] Lior Rokach and Oded Maimon (2005) Top-down induction of decision trees classifiers-a survey, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), pp. 476–487. https://doi.org/10.1109/tsmcc.2004.843247

[22] Rodrigo Coelho Barros, Márcio Porto Basgalupp, Andre CPLF De Carvalho, and Alex A Freitas (2012) A survey of evolutionary algorithms for decision-tree induction, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3), pp. 291–312. https://doi.org/10.1109/tsmcc.2011.2157494

[23] Wei-Yin Loh (2011) Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), pp. 14–23.

[24] Wei-Yin Loh (2014) Fifty years of classification and regression trees, *International Statistical Review*, 82(3), pp. 329–348. https://doi.org/10.1111/insr.12016

[25] Shlomo Geva and Joaquin Sitte (1991). Adaptive nearest neighbor pattern classification, *IEEE Transactions on Neural Networks*, 2(2), pp. 318–322. https://doi.org/10.1109/72.80344

[26] Se June Hong (1997) R-mini: An iterative approach for generating minimal rules from examples. *IEEE Transactions on Knowledge and Data Engineering*, 9(5), pp. 709–717. https://doi.org/10.1109/69.634750

[27] Eric WT Ngai, Li Xiu, and Dorothy CK Chau (2009) Application of data mining techniques in customer relationship management: A literature review and classification, *Expert systems with applications*, 36(2), pp. 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

[28] J Ross Quinlan (1987) Generating production rules from decision trees, In *ijcai*, Citeseer, 87, pp. 304–307.

[29] Ye Ren, Le Zhang, and Ponnuthurai N Suganthan (2016) Ensemble classification and regression-recent developments, applications and future directions, *IEEE Computational Intelligence Magazine*, 11(1), pp. 41–53. https://doi.org/10.1109/mci.2015.2471235

[30] S Rasoul Safavian and David Landgrebe (1991) A survey of decision tree classifier methodology, *IEEE transactions on systems, man, and cybernetics*, 21(3), pp. 660–674. https://doi.org/10.1109/21.97458

[31] Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim (2002) A personalized recommender system based on web usage mining and decision tree induction, *Expert systems with Applications*, 23(3), pp. 329–342. https://doi.org/10.1016/s0957-4174(02)00052-0

[32] Michael Kearns and Yishay Mansour (1999) On the boosting ability of top-down decision tree learning algorithms, *Journal of Computer and System Sciences*, 58(1), pp. 109–128. https://doi.org/10.1006/jcss.1997.1543

[33] Wei-Yin Loh and Yu-Shan Shih (1997) Split selection methods for classification trees. *Statistica sinica*, pp. 815–840.

[34] Avrim L Blum and Pat Langley (1997) Selection of Relevant Features and Examples in Machine Learning, *Artificial intelligence*, 97(1-2). pp. 245–271. https://doi.org/10.1016/s0004-3702(97)00063-5

[35] Robert Messenger and Lewis Mandell (1972) A modal search technique for predictive nominal scale multivariate analysis, *Journal of the American statistical association*, 67(340), pp. 768–772. https://doi.org/10.1080/01621459.1972.10481290

[36] Gordon V Kass (1980) An exploratory technique for investigating large quantities of categorical data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2), pp. 119–127. https://doi.org/10.2307/2986296

[37] Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, and Piotr Duda (2013) Decision trees for mining data streams based on the gaussian approximation, *IEEE Transactions on Knowledge and Data Engineering*, 26(1), pp. 108–119. https://doi.org/10.1109/tkde.2013.34

[38] J. Ross Quinlan (1986) Induction of decision trees, *Machine learning*, 1(1), pp. 81–106.

[39] Salvatore Ruggieri (2002) Efficient C4.5 [classification algorithm], *IEEE transactions on knowledge and data engineering*, 14(2), pp. 438–444.

[40] Leo Breiman, 2017 *Classification and regression trees*, Routledge.

[41] Haidi Rao, Xianzhang Shi, Ahoussou Kouassi Rodrigue, Juanjuan Feng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan, and Lichuan Gu (2019) Feature selection based on artificial bee colony and gradient boosting decision tree, *Applied Soft Computing*, 74, pp. 634–642. https://doi.org/10.1016/j.asoc.2018.10.036

[42] B Chandra and P Paul Varghese (2009) Fuzzifying gini index based decision trees, *Expert Systems with Applications*, 36(4), pp. 8549–8559. https://doi.org/10.1016/j.eswa.2008.10.053

[43] Wei-Yin Loh and Nunta Vanichsetakul (1988) Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association*, 83(403), pp. 715–725. https://doi.org/10.1080/01621459.1988.10478652

[44] Xiao-Bai Li, James R Sweigart, James TC Teng, Joan M Donohue, Lori A Thombs, and S Michael Wang (2003) Multivariate decision trees using linear discriminants and tabu search, *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, 33(2), pp. 194–205. https://doi.org/10.1109/tsmca.2002.806499

[45] Richard J Light and Barry H Margolin (1971) An analysis of variance for categorical data, *Journal of the American Statistical Association*, 66(335), pp. 534–544.

https://doi.org/10.1080/01621459.1971.10482297

[46] Dursun Delen, Cemil Kuzey, and Ali Uyar (2013) Measuring firm performance using financial ratios: A decision tree approach, *Expert Systems with Applications*, 40(10), pp. 3970–3983. https://doi.org/10.1016/j.eswa.2013.01.012

[47] Hyunjoong Kim and Wei-Yin Loh (2001) Classification trees with unbiased multiway splits, *Journal of the American Statistical Association*, 96(454), pp. 589–604. https://doi.org/10.1198/016214501753168271

[48] João Gama (2004) Functional trees, *Machine Learning*, 55(3), pp. 219–250.

[49] Torsten Hothorn, Kurt Hornik, and Achim Zeileis (2015) ctree: Conditional inference trees, *The Comprehensive R Archive Network*, pp. 1–34.

[50] Jianhua Dai and Qing Xu (2013) Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, *Applied Soft Computing*, 13(1), pp. 211–221. https://doi.org/10.1016/j.asoc.2012.07.029

[51] Vadlamani Ravi, H Kurniawan, Peter Nwee Kok Thai, and P Ravi Kumar (2008) Soft computing system for bank performance prediction, *Applied soft computing*, 8(1), pp. 305–315. https://doi.org/10.1016/j.asoc.2007.02.001

[52] Vikas Chaurasia and Saurabh Pal (2013) Early prediction of heart diseases using data mining techniques, *Caribbean Journal of Science and Technology*, 1, pp. 208–217.

[53] Hyunjoong Kim, Wei-Yin Loh, Yu-Shan Shih, and Probal Chaudhuri (2007) Visualizable and interpretable regression models with good prediction power, *IIE Transactions*, 39(6), pp. 565–579. https://doi.org/10.1080/07408170600897502

[54] Gordon V Kass (1975) Significance testing in automatic interaction detection (AID), *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2), pp. 178–189.

[55] Dan Steinberg and Phillip Colla (2009) Cart: classification and regression trees, *The top ten algorithms in data mining*, 9, pp. 179.

[56] Jerome H Friedman (1991) Multivariate Adaptive Regression Splines, *The annals of statistics*, 19(1), pp. 1–67.

[57] Tian-Shyug Lee and I-Fei Chen (2005) A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, *Expert Systems with Applications*, 28(4), pp. 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

[58] Mohammad Rezaie-balf, Sujay Raghavendra Naganna, Alireza Ghaemi, and Paresh Chandra Deka (2017) Wavelet coupled mars and M5 model tree approaches for groundwater level forecasting, *Journal of hydrology*, 553, pp. 356–373. https://doi.org/10.1016/j.jhydrol.2017.08.006

[59] Wei-Yin Loh (2002) Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, pp. 361–386.

[60] John R Quinlan et al (1992) Learning with continuous classes, In *5th Australian joint conference on artificial intelligence*, World Scientific, 92, pp. 343–348.

[61] Behrooz Keshtegar, Cihan Mert, and Ozgur Kisi (2018) Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM, MARS and M5 model tree, *Renewable and Sustainable Energy Reviews*, 81, pp. 330–341. https://doi.org/10.1016/j.rser.2017.07.054

[62] Yong Wang and Ian H Witten (1996) Induction of model trees for predicting continuous classes.

[63] Ali Behnood, Jan Olek, and Michal A Glinicki (2015) Predicting modulus elasticity of recycled aggregate concrete using $M5'$ model tree algorithm, *Construction and Building Materials*, 94, pp. 137–147.https://doi.org/10.1016/j.conbuildmat.2015.06.055

[64] Lisham Bonakdar and Amir Etemad-Shahidi (2011) Predicting wave run-up on rubble-mound structures using M5 model tree, *Ocean Engineering*, 38(1), pp. 111–118. https://doi.org/10.1016/j.oceaneng.2010.09.015

[65] John A Ross and Sook Bang (1996) The AID computer programme, used to predict adoption of family planning in koyang, *Population studies*, 20(1), pp. 61–75. https://doi.org/10.1080/00324728.1966.10406084

[66] Aliakbar Gholampour, Iman Mansouri, Ozgur Kisi, and Togay Ozbakkaloglu (2018) Evaluation of mechanical properties of concretes containing coarse recycled concrete aggregates using multivariate adaptive regression splines (MARS), M5 model tree (M5tree), and least squares support vector regression (LSSVR) models, *Neural Computing and Applications*, pp. 1–14. https://doi.org/10.1007/s00521-018-3630-y

[67] A Etemad-Shahidi and Javad Mahjoobi (2009) Comparison between M5' model tree and neural networks for prediction of significant wave height in lake superior, *Ocean Engineering*, 36(15-16), pp. 1175–1181. https://doi.org/10.1016/j.oceaneng.2009.08.008

[68] Mehrshad Samadi, Ebrahim Jabbari, and H Md Azamathulla (2014) Assessment of M5' model tree and classification and regression trees for prediction of scour depth below free overfall spillways, *Neural Computing and applications*, 24(2), pp. 357–366. https://doi.org/10.1007/s00521-012-1230-9

# Teeth Segmentation of Bitewing X-Ray Images Using Wavelet Transform

Sina Salimzadeh
Department of Electrical & Electronics Engineering, Girne American University
Karaoglanoglu, Kyrenia, Mersin10, Turkey
E-mail: sina.salimzadeh@gmail.com

Sara Kandulu (Izadpanahi)
Faculty of Engineering Technopark Building, Girne American University
University Drive, Karmi Campus, Karaoglanoglu, Mersin10, Turkey
E-mail: sarakandulu@gau.edu.tr

**Keywords:** teeth segmentation, bitewing X-ray images, dental radiographs enhancement, wavelet transform, morphological operations

**Received:** May 12, 2019

*Within the recent twenty years, the dental X-ray images have widely been employed in forensic odontology for human identification, particularly where mass disasters happen. In this paper, a novel method is proposed for the process of teeth segmentation and individual teeth isolation of Bitewing X-ray radiographs. The main objective of this study is to develop an automatic teeth segmentation approach that can be used in an Automated Dental Identification System (ADIS).*

*The proposed method is based on separating teeth according to edge lines between crowns of teeth. It comprises four phases as image enhancement, edge detection by using wavelet transform, Region of Interest (ROI) definition, and morphological processing. Image enhancement in our case is done by image sharpening using a Butterworth high pass filter. Directional changes of the image and a blurred version of it are obtained by wavelet transform in the second phase. In ROI definition the upper and lower jaws are first separated using the integral intensity projection and then a region containing the desired edge lines are defined. In the final stage, some morphological operations are applied to isolate the teeth based on separating edge lines.*

*The evaluation of the teeth segmentation is measured by isolating accuracy and visual inspection. Experimental results with 90.6% isolation accuracy of total 681 teeth illustrate that the proposed method is more efficient that the existing algorithms.*

*Povzetek: Predstavljena je izvirna metoda analize posnetkov zob za namene forenzične identifikacije in verifikacije.*

## 1 Introduction

X-ray radiographs have greatly been used within a variety of medical images. One of the common usages of X-ray imaging is in dentistry and forensic odontology. As tooth is the hardest tissue in our body, it plays an important role in forensic medicine. Individual characteristics such as fingerprints, pupils and face are not always possible for postmortem identification, particularly under the critical circumstances [17]. There are situations such as natural phenomenon (tsunami, hurricane, earthquake, etc.), terrorist attacks, airplane crashes and bomb explosion where victims cannot be identified by visual means. This is where dental features become important for forensic experts [11].

It is widely accepted that image segmentation is the most challenging part of the process of feature extraction from dental X-ray images. In the recent years, several approaches in image segmentation have been introduced and made progress in segmentation in order to overcome the existing shortcomings. The biggest challenges in this process are low quality of X-ray images, noise, and low contrast. However, one of the major problems in dental X-ray images is similarity of the pixel intensities between gum tissue and teeth. Although the inhomogeneity in pixel intensities has several effective factors, the basic problem is the device that produces these radiations [1].

Different methods, including thresholding-based segmentation, edge-based segmentation, clustering-based

Figure 1: Similarity between gum tissue and teeth in a dental X-ray image.

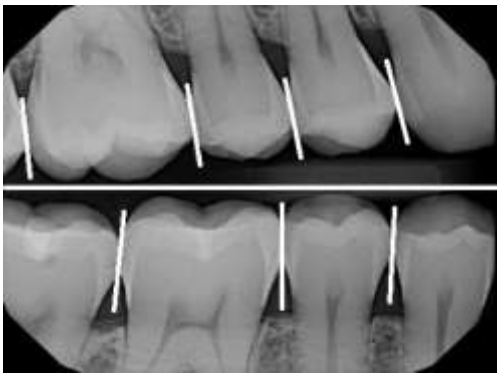


Figure 2: Separating lines for teeth segmentation.

segmentation and region-based segmentation, are being used among automatic processes for teeth segmentation. Thresholding-based techniques rely mainly on the distribution of pixel intensities and they use the information based on the single-band image. Edge-based segmentation methods are prone to produce disjoint edges. Clustering-based methods require training samples; for instance, K-means clustering requires initialization for the number of clusters k. Region-based techniques require objects with similar features in order to segment. These algorithms can segment high contrast simple medical images without noise [15].

The structure of desired segments should be homogeneous. For more complex image segmentation problems, a combination of mentioned segmentation algorithms can be used. Al-sherif, Guo & Ammar [2] use a two-step thresholding technique to binarize the image and then they separate teeth by finding the minimum cumulative energy path. Nomir & Abdel-Mottaleb [3] start segmentation using iterative thresholding followed by adaptive thresholding to segment the teeth from both the background and the bone areas. To separate individual teeth, they use vertical integral protection. Ølberg & Goodwin [4] proposed a path-based technique to segment a dental X-ray image into individual teeth. Abdel-Mottaleb et al. [5] first separate the teeth from the background of the image using a two-step threshold method. In the second stage, they separate each tooth using integral projection.

This study aims to find an automatic approach for teeth segmentation of dental X-ray images with higher accuracy. We propose a novel method that segments the bitewing images according to the edge lines between crowns of the teeth.

# 2 Application of wavelet transform

The proposed method consists of two steps; first, separating upper and lower jaws and second, finding the angle of edge lines between the teeth crowns and using morphological operation to separate each tooth according to the corresponding line.

Several phases are required to convert a dental X-ray image to the extracted features of each tooth as follows:

- Preprocessing
- Edge detection
- ROI definition

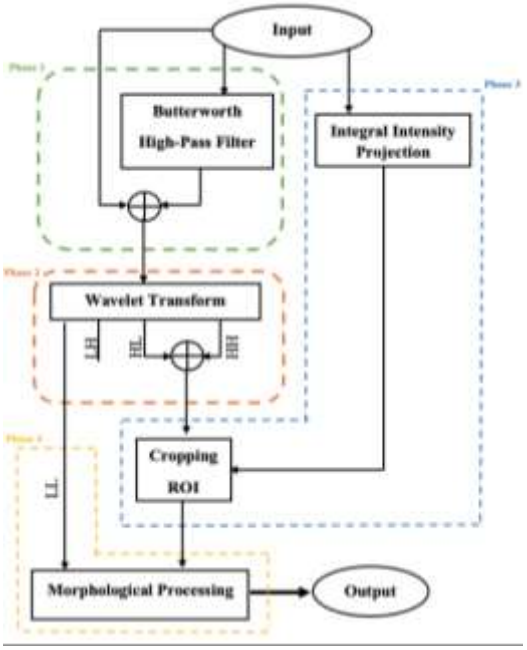



Figure 3: Block diagram of the proposed method.

- Teeth separation

## 2.1 Preprocessing

The nature of intra-oral radiographs includes poor image quality and so an efficient technique for image enhancement is required. Enhancement is an essential step in preprocessing of X-ray images due to poor quality, noise, unclear region boundaries [6]. Image enhancement techniques are applied in the new method proposed in this paper to reduce noise and increase contrast between different layers of the image [7]. Several efficient sharpening techniques have been employed to enhance dental X-ray images. The newly introduced method is done among homomorphic High Pass Filter (HPF), morphological top hat and bottom hat filter, Butterworth HPF and Gaussian HPF. The results can be seen in Figure 4.

It can be seen that the sharpened image by Butterworth HPF has lighter teeth and darker gums. The advantages of Butterworth HPF have been considered in this method by employing a general transfer function as follows:

$$H(u, v) = \frac{1}{1+[\frac{D_0}{D(u,v)}]^{2n}}$$

where $D_0$ is the cut-off frequency, n is the order of the filter and D(u, v) is the distance between a point (u, v) in the frequency domain and the center of frequency plane.

It has been found that the sharpened image contains high contrast between gap regions and the teeth that makes it easier to find the edge lines.

## 2.2 Edge detection

Edge detection is one of the effective methods to perform image segmentation. There are various types of edge detector operators such as Canny operator, Sobel operator, Prewitt operator, Robert's operator and LoG operator. Various disadvantages have been found about these

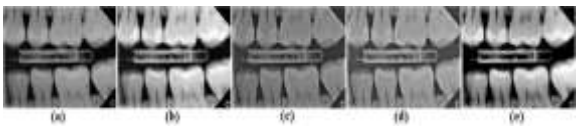


Figure 4: Comparison of sharpening methods. (a) input image, (b) morphological top hat / bottom hat, (c) homomorphic HPF, (d) Gaussian HPF, (e) Butterworth HPF.

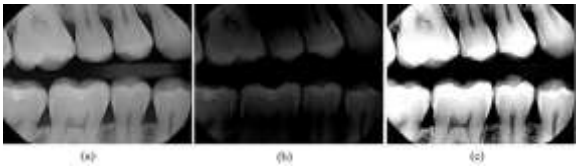


Figure 5: Dental image enhancement. (a) input image, (b) Filtered image, (c) Sharpened image.

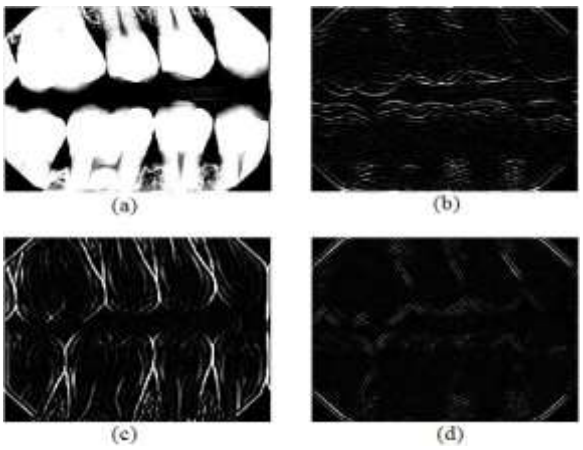



Figure 6: Bitewing X-ray image after decomposition by wavelet transform. (a) approximation, (b) horizontal details, (c) vertical details, (d) diagonal details.

methods as difficulties with detection of minor details, and a need to have high quality images to perform the edge detection operators in a satisfiable level [4]. Thus, in noisy images or low-quality images, these methods are not able to distinguish between edges and noise components.

A technique for edge detection that can overcome the mentioned problem is to use discrete wavelet transform (DWT) [8]. Down sampling in each sub-band of DWT leads to information loss in the output image. Therefore, Stationary Wavelet Transform (SWT) has been employed in the proposed method to overcome this loss.

The process of applying SWT to an image can be represented as a set of filters [9]. As shown in Figure 6, the image is divided into four bands including LL (approximation), LH (horizontal details), HL (vertical details), and HH (diagonal details). The letters H and L represent High pass and Low pass filtering respectively in each stage.

Among these three detailed images (horizontal, vertical and diagonal) we only need two bands of vertical and diagonal details in order to separate teeth in each jaw. A simple solution for this challenge is to combine vertical and diagonal details.

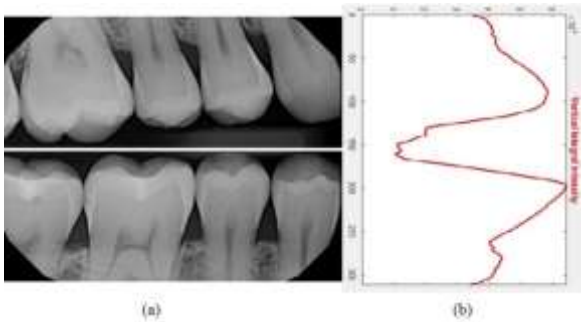


Figure 7: Gap finding between upper and lower jaw. (a) input image, (b) vertical intensity projection.

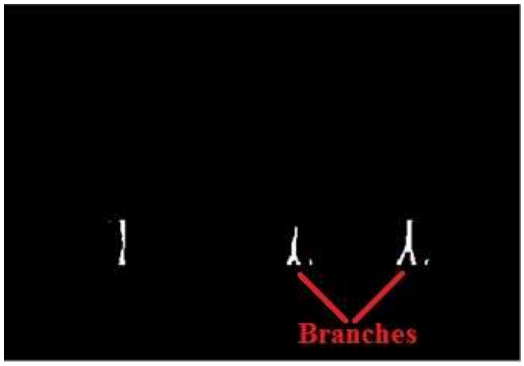



Figure 8: Branches of the edge lines in lower jaw.

## 2.3 ROI definition

The first step in region of interest (ROI) definition phase is jaw separation by finding the gap between jaws. Vertical intensity projection has been used for jaw separation. For the given image with intensity function I(x, y), the vertical intensity projection is defined as follows:

$$V(x) = \sum_{y=y_1}^{y_2} I(x, y)$$

The second step in ROI definition is cropping the region that contains crowns. Once cropped, details in the region containing gum tissue and roots cannot be seen anymore.

## 2.4 Morphological processing

After separating the image (containing edge lines) into two upper and lower jaws, it is needed to break branches of the existing edge lines. For this purpose, the morphological operations are used to skeletonize the edge lines and break all the branches, because only those parts of edge lines are of interest where crowns are connected.

After breaking the branches, small objects and noise in the binary image can be removed and the orientation of the separating edge lines in both upper and lower jaws can be found. By knowing the orientation of the separating lines, the desired structuring element can be defined and then the morphological image opening can be applied to separate teeth from each other.

Opening an image A by a structuring element B is denoted by $A \circ B$ and is defined as below:

$$A \circ B = (A \ominus B) \oplus B$$

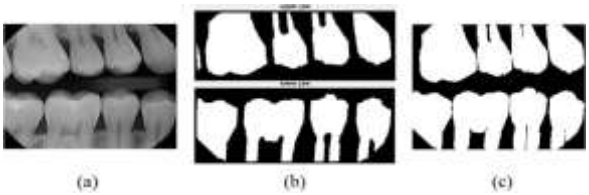


Figure 9: Morphological processing. (a) input image, (b) the opened binary images of upper/lower jaw, (c) the thickened image of the stitched image.

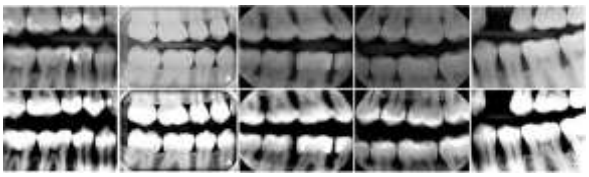


Figure 10: Upper row: original input images; lower row: sharpened Images by Butterworth HPF.

Morphological opening removes the regions of an object that cannot contain the structuring element. It breaks thin connections, smooths object contours and removes thin protrusions [10]. This leads to separated teeth in both upper and lower jaws. Finally, morphological image thickening is applied to retrieve size decrement due to image opening.

# 3 Results and discussion

## 3.1 Experimental results

First of all, the dental X-ray radiographs are required to get enhanced due to low quality and contrast. Different techniques have been used for this purpose by different authors; for instance, the morphological top-hat and bottom-hat filtering is used to increase contrast in medical images [11]; [7]; [6]; [12]; [13]; [14]; [15]; [16]; [4]. The histogram equalization as well as a combination of homomorphic and Butterworth HPF for image enhancement are employed for this purpose [17]; [18]; [19]; [20].

To decide which type of filtering results better, a comparison is done among four well-known types of medical images filtering. Table 1 shows the comparison of four bitewing X-ray images based on Peak Signal-to-Noise Ratio (PSNR) value of the input image and the sharpened image.

As can be seen in Figure 10, by applying Butterworth HPF, the teeth parts become lighter and the gum parts become darker that means the increment of contrast in the image.

Expanding the histogram of the image by sharpening strengthens the differences between various tissues in dental X-ray images. This prepares the image for finding directional details in the next phase. The gap between upper and lower jaw creates a valley in the graph of the vertical integral intensity projection. Hence, it can be concluded that this is a suitable approach to separate jaws.

The next step in ROI definition is to remove the undesired details such as roots and top of the teeth. Morphological operators are applied to obtain separating lines. Then, the binary version of approximation image is opened by separating edge lines. In the final stage, the location of individual teeth and extract them has been found by labeling connected components.

| PSNR (dB) | | | | |
|---|---|---|---|---|
| Filtering Type |  |  |  |  |
| Butterworth HPF | 35.23 | 36.27 | 36.73 | 35.77 |
| Morphological Top/ Bottom-hat | 31.07 | 34.06 | 33.32 | 33.42 |
| Gaussian HPF | 34.56 | 35.92 | 36.3 | 35.66 |
| Homomorphic HPF | 34.27 | 35.74 | 35.66 | 35.2 |

Table 1: PSNR of filtering in dental X-ray images.

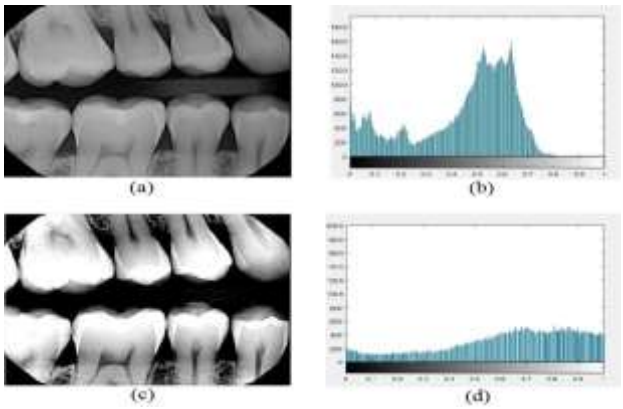


Figure 11: Contrast increment. (a) input image, (b) histogram of the input image, (c) sharpened image, (d) histogram of the sharpened image.

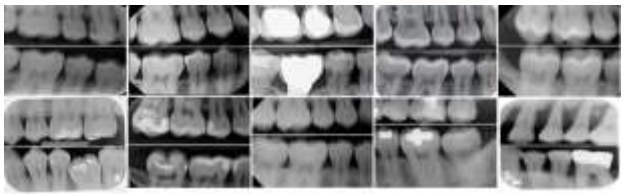


Figure 12: Some examples of jaws separation.

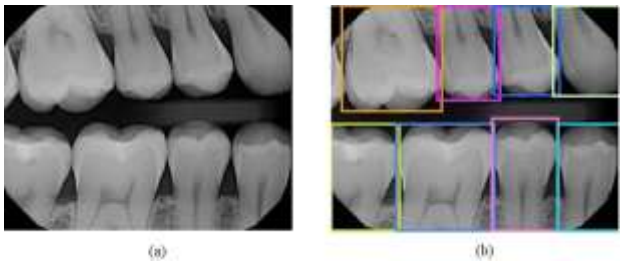


Figure 13: Teeth detection. (a) original input image, (b) bounding box of each tooth.

## 3.2 Evaluation

The algorithm is implemented in MATLAB R2016b, using an Intel(R) Core(TM) i5 CPU at 1.70GHz and 2.40GHz with 4GB RAM in a Microsoft Windows 8.1 Pro environment.

| Segmentation method | Correctly separated in upper jaw | Correctly separated in lower jaw | Total isolation accuracy |
|---|---|---|---|
| Abdel-Mottaleb et al. [5] | 169/195 – 86.66% | 149/181 – 82.32% | 318/376 – 84.57% |
| Nomir & Abdel-Mottaleb [3] | 329/391 – 84.14% | 293/361 – 81.16% | 622/752 – 82.71% |
| Al-Sherif et al. [2] | 1604/1833 – 87.5% | 1422/1692 – 84% | 3026/3525 – 85.8% |
| Ølberg & Goodwin [4] | 300/336 – 89.3% | 270/306 – 88.2% | 570/642 – 88.78% |
| The proposed method | 325/351 – 92.6% | 292/330 – 88.5% | 617/681 – 90.6% |

Table 2: The result of teeth segmentation methods.

| Segmentation method | Correctly separated in upper jaw | Correctly separated in lower jaw | Total isolation accuracy |
|---|---|---|---|
| Ølberg & Goodwin [4] | 308/351 – 87.7% | 285/330 – 86.3% | 593/681 – 87% |
| The proposed method | 325/351 – 92.6% | 292/330 – 88.5% | 617/681 – 90.6% |

Table 3: The comparison using the same database.

85 bitewing X-ray images have been used for teeth segmentation experiments with the total 681 separable teeth. Teeth are divided into two groups as the teeth in the upper jaw and the teeth in the lower jaw. The evaluation of segmentation is based on the isolation accuracy.

$$\text{Isolation Accuracy} = \frac{N_c}{N_t} \times 100\%$$

where Nc is the number of teeth that are correctly isolated and Nt is the total separable teeth.

In the proposed method, 325 teeth in the upper jaw and 292 teeth in the lower jaw are separated correctly out of the total 351 and 330 teeth in the upper and lower jaws, respectively.

Among plenty of different approaches, Table 2 shows the isolation accuracy of four efficient methods for upper/lower jaw and overall teeth separately. It has been shown that the proposed method has the highest performance in correctly separating the teeth at both upper and lower jaws. Consequently, it attains the best isolation accuracy among the other state of the art methods.

The proposed method has also been compared with Ølberg & Goodwin [4], with the best performance in comparison with the other state of the art method, using the same database containing 681 separable teeth and the result of this comparison is shown in Table 3.

Ølberg & Goodwin in [4] use morphological top-hat and bottom-hat filtering to enhance the input image and then they separate teeth by using path-based method. It can be seen that the proposed method achieves better result and can detect and separate the teeth in upper and lower jaws with higher accuracy.

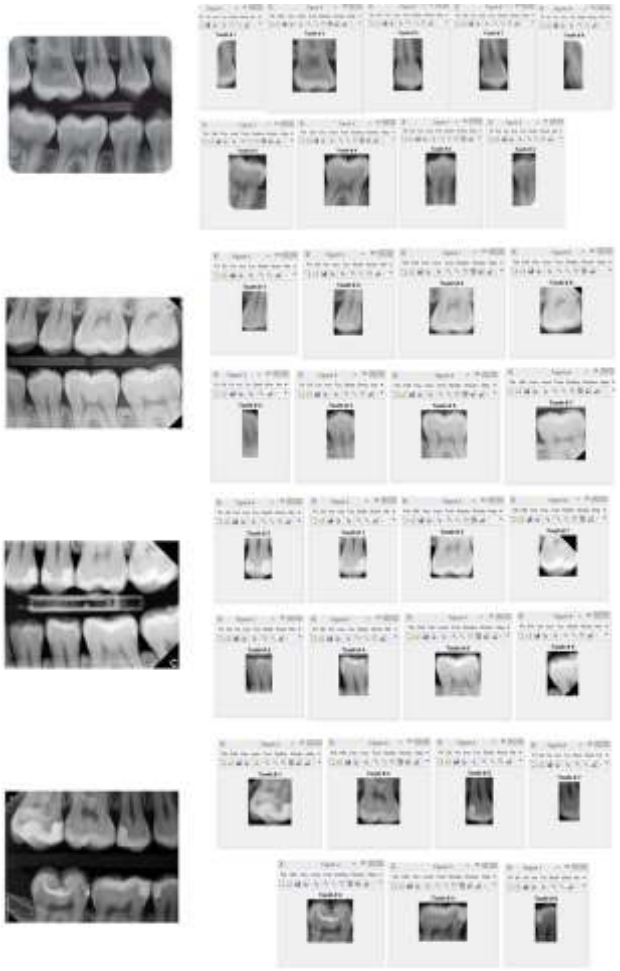


Figure 14: Some examples of the final results.

# 4 Conclusions

Poor quality of the X-ray images prevents the teeth separation and causes under- or over-segmentation. To solve this problem, a novel method for separating the bitewing X-ray image into individual teeth has been introduced in this paper. At the first phase of the proposed method, the resolution of the image is enhanced and afterwards a wavelet-based edge detection followed by some morphological operations segments and separates each tooth.

The experimental result of 90.6% in terms of isolation accuracy on a database consisting of 681 teeth in 85 bitewing X-ray images has been employed in this paper to show the superiority of the proposed method in comparison to the state-of-the-art teeth segmentation methods. The method can be used as a part of an ADIS for matching purpose.

# 5 References

[1] Chunming Li, Rui Huang, Zhaohua Ding, Chris Gatenby, Dimitris Metaxas, and John Gore. A variational level set approach to segmentation and bias correction of images with intensity inhomogeneity. In: Metaxas D., Axel L., Fichtinger G., Székely G. (eds) Medical image computing and

computer-assisted intervention. *MICCAI 2008*. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 5242:1083-1091, 2008. https://doi.org/10.1007/978-3-540-85990-1_130.

[2] Nourdin Al-sherif, Guodong Guo, and Hany H. Ammar. A new approach to teeth segmentation. *2012 IEEE International Symposium on Multimedia*, Irvine, CA, 145-148, 2012. https://doi.org/10.1109/ism.2012.35.

[3] Omaima Nomir, and Mohamed Abdel-Mottaleb. A system for human identification from X-ray dental radiographs. *Pattern Recognition*, Elsevier, 38(8):1295-1305, 2005. https://doi.org/10.1016/j.patcog.2004.12.010.

[4] Jan-Vidar Ølberg, and Morten Goodwin. Automated dental identification with lowest cost path-based teeth and jaw separation. *Scandinavian Journal of Forensic Science*, Sciendo, 22(2):44-56, 2016. https://doi.org/10.1515/sjfs-2016-0008.

[5] Mohamed Abdel-Mottaleb, Omaima Nomir, Diaa Eldin Nassar, Gamal Fahmy, and Hossam Hassan Ammar. Challenges of developing an automated dental identification system. *46th Midwest Symposium on Circuits and Systems*, IEEE, Cairo, 1:411-414, 2003. https://doi.org/10.1109/mwscas.2003.1562306.

[6] Jindan Zhou, and Mohamed Abdel-Mottaleb. A content-based system for human identification based on bitewing dental X-ray images. *Pattern Recognition*, Elsevier, 38(11):2132-2142, 2005. https://doi.org/10.1016/j.patcog.2005.01.011.

[7] Gamal Fahmy, Diaa Nassar, Eyad Haj-Said, Hong Chen, Omaima Nomir, Jindan Zhou, Robert Howell, Hany H. Ammar, Mohamed Abdel-Mottaleb, Anil K. Jain. Towards an automated dental identification system (ADIS). In: Zhang D., Jain A.K. (eds) Biometric authentication. *ICBA 2004*. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 3072:789-796, 2004. https://doi.org/10.1007/978-3-540-25948-0_107.

[8] Darshan Bhavesh Mehta, and Harsha Kosta. Image compression using discrete cosine transform. *International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)*, 4(4):6796-6799, 2016. DOI: 10.15680/IJIRCCE.2016. 0404073.

[9] Kamlesh Kumar, Nadir Mustafa, Jian-Ping Li, Riaz Ahmed Shaikh, Saeed Ahmed Khan, and Asif Khan. Image edge detection scheme using wavelet transform. *11th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, 261-265, 2014. https://doi.org/10.1109/iccwamtip.2014.7073404.

[10] Rafael C. Gonzalez, and Richard E. Woods. *Digital image processing, 2nd Edition*. Pearson. 2002.

[11] Eyad Haj Said, Diaa Eldin Nassar, Gamal Fahmy, and Hany H. Ammar. Teeth segmentation in digitized dental X-ray films using mathematical morphology. *IEEE Transactions on Information Forensics and Security*, 1(2):178-189, 2006. https://doi.org/10.1109/tifs.2006.873606.

[12] Mohammad H. Mahoor, and Mohamed Abdel-Mottaleb. Automatic classification of teeth in bitewing dental images. *International Conference on Image Processing*, ICIP '04., Singapore, 5:3475-3478, 2004. https://doi.org/10.1109/icip.2004.1421863.

[13] Eyad Haj Said, Gamal Fahmy, Diaa Eldin Nassar, and Hany H. Ammar. Dental x-ray image segmentation. *Proceedings of SPIE - The International Society for Optical Engineering*, United States, 5404:409-418, 2004. https://doi.org/10.1117/12.541658.

[14] Jindan Zhou, and Mohamed Abdel-Mottaleb. Automated human identification based on dental X-ray images. *Proceedings of Biometric Technology for Human Identification*, United States, 5404:373-380, 2004. https://doi.org/10.1117/12.542689.

[15] Mohammad H. Mahoor, and Mohamed Abdel-Mottaleb. Classification and numbering of teeth in dental bitewing images. *Pattern Recognition*, Elsevier, 38(4):577-586, 2005. https://doi.org/10.1016/j.patcog.2004.08.012.

[16] Faraein Aeini, and Fariborz Mahmoudi. Classification and numbering of posterior teeth in bitewing dental images. *3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, IEEE, Chengdu, V6-66-V6-72, 2010. https://doi.org/10.1109/icacte.2010.5579369.

[17] Noorhayati Mohamed Noor, Noor Elaiza Abdul Khalid, Mohd Hanafi Ali, and Alice Demi Anak Numpang. Enhancement of soft tissue lateral neck radiograph with fish bone impaction using adaptive histogram equalization (AHE). *Second International Conference on Computer Research and Development*, IEEE, Kuala Lumpur, 163-167, 2010. https://doi.org/10.1109/iccrd.2010.84.

[18] Thangavel Kuttiyannan, R. Manavalan, and Laurence Aroquiaraj. Removal of speckle noise from ultrasound medical image based on special filters: comparative study. *ICGST-GVIP Journal*, 9(3):25-32, 2009.

[19] Peyman Rahmati, Ghassan Hamarneh, Doron Nussbaum, and Andy Adler. A new preprocessing filter for digital mammograms. In: Elmoataz A., Lezoray O., Nouboud F., Mammass D., Meunier J. (eds) Image and signal processing. *ICISP 2010*. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 6134:585-592, 2010. https://doi.org/10.1007/978-3-642-13681-8_68.

[20] Vijayakumari Pushparaj, Ulaganathan Gurunathan, and Banumathi Arumugam. An effective dental shape extraction algorithm using contour information and matching by mahalanobis distance. *Journal of Digit Imaging*, Springer, 26(2):259–268 2013. https://doi.org/10.1007/s10278-012-9492-4.

# E-learning in Business Practice, a Case Study During COVID-19 in Croatia

Dominika Crnjac Milić, Zdravko Krpić and Filip Sušac
J. J. Strossmayer University of Osijek, Faculty of Electrical Engineering
Computer Science and Information Technology Osijek, Kneza Trpimira 2b, 31000 Osijek, Croatia
E-mail: dominika.crnjac@ferit.hr, zdravko.krpic@ferit.hr, filip.susac@ferit.hr

**Keywords:** business practice, COVID-19, Croatia, e-learning, employee education

**Received:** July 31, 2020

*Technological progress is increasingly changing the form of learning and communication in the business environment. The omnipresence of computers enables distance learning to an increasing number of people, whose systematic use is enabled primarily via various e-learning platforms. To be a market competitive, companies must be ready for fast changes and constant adaption to the new technologies. This paper provides a brief classification of e-learning its features and its content as well as an overview of the presence of e-learning in the companies from the Republic of Croatia via a survey that we conducted among 80 of them. The results of the survey are particularly important since they were collected during the COVID-19 pandemic when companies were abruptly forced to adapt to working from home. The primary goal of the survey was to analyze to which extent they use e-learning for their daily business and education of the employees. The survey also shows what services do they use and prefer the most and what is their overall opinion about e-learning in business.*

*Povzetek: V prispevku je analizirano e-učenje v 80 hrvaških podjetjih za časa krize COVID-19.*

## 1 Introduction

Nowadays, running a competitive business organization considerably depends on the ability to adapt to the new technologies and working conditions. Investing in employees in terms of learning and mastering new skills results in greater productivity and better time management [1]. This is especially evident during crises, such as an ongoing COVID-19 pandemic, which is one of the biggest that humanity saw in recent history [2].

In this paper, we aim to capture aspects of e-learning in the business sector, such as the expected benefits from it, frequency of its use, technologies utilized, and the general opinion of its users and company managers. These aspects are rather important considering that they were collected during the COVID-19 pandemic, which acts as a stress test for the companies' ability to embrace e-learning in their businesses. We collected the most important information about the use of e-learning and investing in employee education, via a survey conducted among 80 Croatian companies.

There are numerous advantages of utilizing e-learning in business practice, starting from the possibility of educating employees at any place at any time, intensified mutual interaction, and significant savings in time and money invested in employee education [3]. Via a survey, we analyzed if the Croatian companies recognize the importance of e-learning and whether they utilize its full potential.

One set of questions within the survey was designed specifically to evaluate the importance and frequency of use for various types of content present within the e-learning platforms that participants use, as well as to identify the challenges that may arise during their use. To put the collected answers into context, we provided an overview of the advantages and disadvantages of e-learning beforehand, as well as their classification. This way we give a good estimate of e-learning options available to the users today. Thus, this paper aimed to provide theoretical and empirical background on e-learning and verify it by the Croatian business practice use case.

This paper is organized as follows. In Section 2., a general overview and classification of the e-learning is presented. Section 3. describes the survey conducted among companies in Croatia, whereas Section 4. provides the analysis of the survey. Section 5. concludes the paper with recommendations for the inclusion of e-learning in businesses.

## 2 The overview of e-learning

Planning, design, development, implementation, and control of an e-learning system are a serious undertaking, regardless of the degree of knowledge of all methodological and technological aspects. Respecting the fact that technologies are developing at astonishing speed and that the access to employee education is often exposed to changes, there is a great potential for failures that needs to be minimized, such as dropouts and insufficient technical knowledge [4]. According to [5], today's e-learning systems can be classified into six main types, as shown in Table 1. In practice, there is a rising trend on

hybrid teaching and online education, which are a part of distance learning. E-learning continuum is given in Figure 1 [6].

E-learning has several definitions, but the most general one is that e-learning is a learning method that uses

| Type of e-learning | Description |
|---|---|
| face-to-face, F2F | e-learning with physical presence and without e-communication |
| self-learning | e-learning without presence and without e-communication |
| asynchronous | e-learning without presence and with e-communication |
| synchronous | e-learning with virtual presence and with e-communication |
| blended/hybrid-asynchronous | e-learning with occasional presence and with e-communication |
| blended/hybrid-synchronous | e-learning with presence and with e-communication |

Table 1: Types of e-learning.

ICT to ubiquitously transform and support teaching and learning process [7]. According to [8], in e-learning participants receive knowledge via the Internet without having to physically be present where the education takes place. Recently, multimedia is progressively more used to improve learning, with a high emphasis on IoT systems.

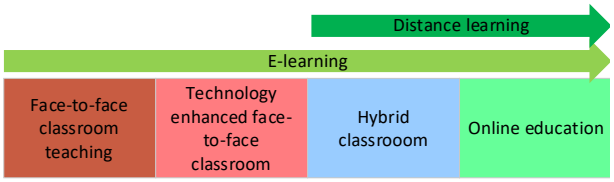



Figure 1: Continuum of e-learning.

In the last decade, the e-learning platforms proliferated due to the increased interest in e-learning. E-learning platforms are web sites that encompass diverse content and high-quality learning opportunities. They enable communication between lecturers and participants, as well as between the participants (or lecturers) themselves. According to [7], e-learning platforms include visualization, animation, simulations, various interactive elements, tests, quizzes, forums, distribution lists, and other types of content. Therefore, preparation for the online execution of one teaching unit is usually a lot more challenging than for the off-line execution. One set of questions within the survey was designed specifically to evaluate the importance and frequency of use for various types of content present within the e-learning platforms that participants use, as well as to identify the challenges that may arise during their use. In order to put the collected answers into context, we provided an overview of the advantages and disadvantages of e-learning beforehand, as well as their classification. This way we give a good estimate of e-learning options available to the users today. Thus, the aim of this paper was to provide theoretical and empirical background on e-learning and verify it by the Croatian business practice use case.

## 2.1 Advantages and disadvantages of the e-learning

There are several reasons why it may be justifiable to develop an online education system for different kinds of courses [9]. For academic purposes, these reasons are the ability to create an online school subject, a college course, or some other kind of a course, as well it contributes to the development of the basic skills needed for lifelong learning [10]. For companies, e-learning is a way to provide lifelong learning, to enable working from a distant location or to upgrade employee skills required for daily tasks. According to [11], users are learning faster by e-learning than by F2F learning, thus making e-learning as a possible answer to the increasingly challenging demands of the business market and society. Some of the identified benefits of e-learning are:

- Possibility of learning anywhere
- Possibility of learning anytime
- Mutual synergy between lectures and participants
- Participants are in the center of attention
- No discrimination
- Creative learning
- Greater access to resources
- Simplified feedback
- Tailoring one's own education

However, e-learning has also some negative aspects that are still not completely remedied in practice and therefore can reduce productivity and the amount of knowledge transferred. One of them is considered to be the lack of communication between the participants, which was shown to be the main reason for the participant dropout, [12]. In [13] the authors have shown that most respondents believe e-learning can help in developing new skills. But most of them, at the same time, feel that they have insufficient knowledge and experience to take a full advantage of the benefits of e-learning. Additionally, they believe that newer generations are more likely to accept e-learning, considering that they grew up with computers and mobile phones, unlike older generations. This realization points to the need that the e-learning content in the business sector must be more contextual and dynamic than in a classical academic or school environment. It is also necessary to encourage interaction among the participants of e-learning so that they do not feel isolated, and to be additionally motivated by receiving feedback from other participants on the acquired knowledge. Some other identified disadvantages of e-learning are:

- Obligatory access to the computer
- The necessary basic knowledge of using computer
- Insecurity of technology
- Social isolation
- Creating technology addiction
- Lack of self-discipline of education participants
- Too many participants to achieve optimal synergy
- Not all content is suitable for online education

- The need to maintain and introduce innovations due to new technologies and modern content in distance education
- Inability to transform the traditional curriculum into online education

Another view on the advantages and disadvantages of e-learning, given from the medical perspective is available in [14]. Some of the advantages and disadvantages emerge also from the results of the survey given in Section 4.

## 2.2 E-learning classification

According to [15], e-learning can be classified according to:

- Type of the platform on which it is taking place
- Application area
- Standard
- Type of learning
- Type of tools that the platform can have
- Type of activity

The list of e-learning types for each classification is given in Figure 2.

### 2.2.1 Classification according to the type of platform

Considering the type of the e-learning platform, the three types exist:

- Learning Management System (LMS) – most often used in both the business sector and in higher education. It is a platform that allows one to store and deliver learning content and to monitor the user's participation. The main goal of LMS is participant management and monitoring their activity [9]. In essence, LMS is an automated software for tracking, reporting, administration, registration, and evaluation [16].
- Learning Content Management System (LCMS) – primarily used for creating, storing, and organizing content. Unlike LMS, it is used to create learning content and publish it in various forms. It can be perceived as an upgraded version of LMS with some content management features [9].
- Content Management System (CMS) – unlike LMS and LCMS, it is based mostly on content, and its quick and efficient use. CMSs enable creating online courses, uploading documents and presentations in various formats and other content features [16]. CMSs specialize in the creation and management of learning content.

### 2.2.2 Classification according to the application area

Classification according to the application area divides the platforms into two groups:

- The business sector – the aim is to enable employees to acquire knowledge and skills to increase their competence for work and for executing the specific tasks that are put before them. The main features of the e-learning in the business sector are a fast pace, heavy focus on the task for which education is intended, and the emphasis on achieving maximum results.
- Academic institutions – unlike in the business sector, the main goal is to transfer different aspects of knowledge rather than to train participants to perform certain tasks. According to [17] 99% of the high education institutions have an LMS in place, where 85% of them have been utilized to support their educational services. Although the primary task of e-learning in the academic sector is to improve the educational process, students also gain overall knowledge about e-learning which then serves as a base for further improvement of knowledge and skills.

### 2.2.3 Classification according to standard

Two types of e-learning platforms exist according to e-learning standards:

- Sharable Content Object Reference Mode (SCORM) – an older standard that is slowly being

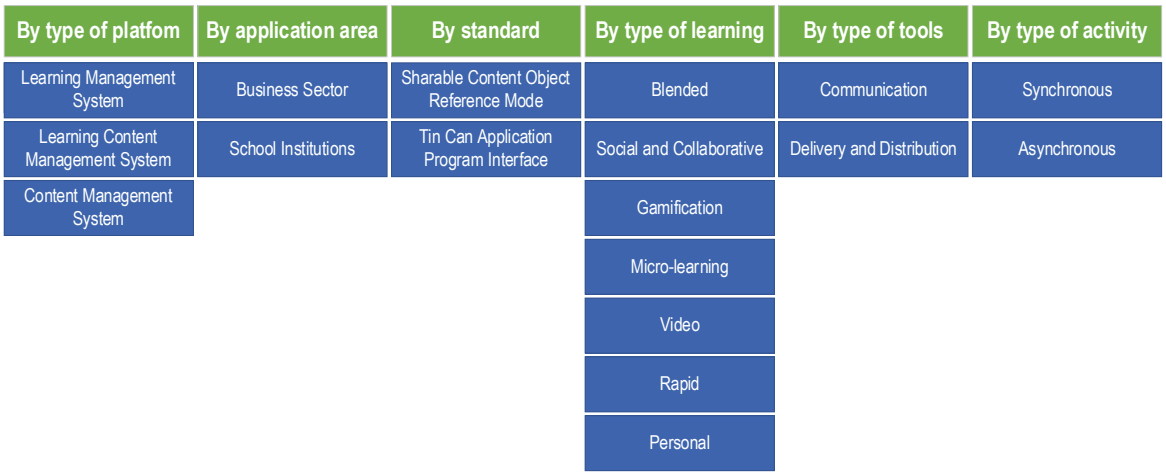



Figure 2: Classification of e-learning.

replaced by Tin Can API. Both standards allow users to run the course, take quizzes, monitor user's work. Still, SCORM is simpler and has fewer options.

- Tin Can Application Program Interface (Tin Can API) – although similar to the SCORM standard, is increasingly being used since it is more reliable, it makes it easier to handle a large number of data, it enables better tracking of the user's work and the most important part is that it is still being upgraded and will be even more advanced in the future.

#### 2.2.4 Classification according to the type of learning

Classification according to the type of learning split e-learning platforms into seven main groups:

- Blended learning – a combination of F2F learning and online learning in a way that the one complements the other. It is also called hybrid learning.
- Social and collaborative learning – learners work together to expand their knowledge of a subject or skill. It is typically done through live chats, message boards, or instant messaging.
- Gamification – the use of game-based mechanics, and game thinking to engage people, promote learning, and solve problems. Games are created to draw people in, to keep them playing, to keep them interested and involved.
- Micro-learning approach – can provide educational benefits without overwhelming the learner. It is quickly becoming one of the most popular e-learning trends.
- Video learning – brings a whole new dimension to the teaching methods because of the theory that everything that is being taught can be demonstrated. Video also helps to add a feeling of personalization to a course. A video of the tutor giving a lecture helps the students to feel a connection.
- Rapid e-learning – essentially a faster process of designing and developing online-based learning courses. Rather than spending months, rapid e-learning allows creators to build lessons and content in a matter of days or weeks. Typically, this is done through PowerPoint or narrated videos and after that, a software is utilized to evaluate the students, as well as to provide them with activities that they can perform on their own in between presentations or videos.
- Personalized e-learning – enables participants to customize a variety of the elements involved in the online education process. This means that they are asked to set their own goals, go at their own pace, and communicate with instructors and participants to personalize the learning process. The most important thing is the feedback that improves learning results.

#### 2.2.5 Classification according to the tools that a platform can have

This classification identifies two main groups of e-learning platforms:

- Communication tools – which include blogs, E-mail, instant messaging, online groups, chats, forums, and web-conferencing.
- Delivery and distribution tools – which include websites, sharing files, and streaming.

#### 2.2.6 Classification according to the type of activity during e-learning

Classification according to the type of activity performed during e-learnings results in two main groups of e-learning platforms:

- Synchronous activities – real-time activities such as web conferencing, instant messaging, and chats. The problem with synchronous activities is that the activities are not time-flexible, and the user must be online at that time when they are running.
- Asynchronous activities – education can be carried out even when a user or a lecturer is offline. The user can attend the education at his own pace. Examples of asynchronous activities are blogs, forums, E-mail.

The types of e-learning enumerated in this chapter show that users have a wide range of knowledge delivery methods available to them. Companies typically use the types of e-learning that offer specific knowledge of a particular area, the fastest delivery of that knowledge and that cost the least.

## 3 Data collection process and survey hypotheses

### 3.1 Research methods and limitations

To investigate the impact, importance, and presence of e-learning in Croatian business practice, we conducted an online survey. The survey encompassed several types of questions. The survey was based on Google Forms and consisted of 26 questions. The types of questions included 15 questions with a single answer, 2 short text questions, and 9 multiple answer questions where participants could range their responses according to the Likert scale. One question was linked with the answer of a previous one, so not all participants had the same number of questions. The survey was distributed among participants using a direct link which was forwarded to the e-mail addresses of the targeted companies. Data were collected in a period of almost two months, from March 20$^{th}$ until May 17$^{th}$ 2020. The target groups of this research (and survey) were company owners, CEOs, and managers as they have the most complete information about the company management.

A total of 80 companies registered in the Republic of Croatia participated in the survey. Some of the companies

that participated in the survey are multinational, and although they are registered in the Republic of Croatia, they operate in synergy with the companies having the same owner in other countries. In addition to having different positions in the company, the participants also have different amount of work experience and were of a different gender.

The main questions of the survey were:

- How much does the use of e-learning platforms contribute to your business?
- Assess the ability of your company's employees to acquire knowledge through e-learning platforms?
- How much knowledge can be acquired through e-learning platforms compared to the traditional teaching methods?
- Do you think your employees respond positively to the obligation to use e-learning platforms?
- How important is a particular form of content for the acquisition of knowledge through e-learning platforms?

The rest of the questions were used to deepen the information collected by the main questions.

This study has potential limitations. The results were obtained on a relatively small sample of companies that were willing to participate. However, we believe that the number of participating companies is big enough to give an insight into the validity of the hypotheses made within this work (due to their nature). Also, it would require a significant effort to increase the sample size considering their busyness and availability. Care was taken to ensure having a sample heterogeneous enough with regards to both size and the type of company's activity, but the results are a bit biased towards big companies. Additionally, there was no equal sampling from different economic activities, which is going to be a part of the future work.

### 3.2 Research goals

The survey was conducted throughout a two months period, during the special working conditions caused by the COVID-19 pandemic, which had a tremendous impact on incorporating new business solutions for business owners and employees in management positions. Uncertainty and increased responsibility imposed by the situation in which they found themselves, affected the need for educating the employees, the need for changes in the organization and mode of operation of the company. The realization of the necessity of using ICT in everyday business has gained another dimension. Many companies have switched their business and communication to an online form. There was also a need to educate employees online. In addition to all the above, a motivation for this research emerged with a primary goal to find out the attitude of the target group of participants towards the acquisition of new knowledge and skills of their employees via e-learning. It is important to emphasize that the participants' position and the responsibility entrusted to them in companies imply that they should be the motivators for their employees to accept new models of education (which also makes significant savings to their companies). Participant opinion about the general acceptance of the online education in their company, and what content they think is relevant for improving the productivity of their employees is one of the key information for creating a e-learning system tailored to the workers need.

According to [18] many e-learning courses are meaningless from the perspective of meeting the education goals. For example, some business organizations only want to meet some of the regulatory guidelines and show that they have provided training that is provided to employees according to standards and trends. Additionally, it has been proved useful to know the general opinion about e-learning of a selected group of participants who create added value daily through their business in the Republic of Croatia, but also abroad.

The hypotheses of our research conducted by this paper were as follows:

**H1**: E-learning programs in business organizations are relatively new and there is a possibility of facing increased caution and distrust of managers and employees towards the manner of e-learning, as well as the results achieved by it.

**H2**: By recognizing the usefulness of e-learning, companies regardless of their size implement them.

**H3**: E-learning is conducted in a targeted and segmented manner, with the aim of raising the overall work performance.

**H4**: E-learning is rarely perceived as an unnecessary burden on top of the already busy working hours, making employees aversive towards it.

## 4 Survey analysis

Great anticipation was linked with survey analysis, since they were collected during the COVID-19 pandemic, thus making them particularly relevant. Namely, during the pandemic, the e-learning platforms had peak use, which exposed all the potential drawbacks and advantages that they may have. During these times, many employees were sent to work from home, as well as were instructed to attend various online training. Also, some companies tried to turn the situation in their favor by using the idle times of their employees to enroll them into various educations, as well as to reorganize their businesses into working from home.

A total of 80 companies participated in the survey, making the results even more relevant for understanding the e-learning practices in the business sector. It is worth to mention that the anonymity of each participant was respected.

### 4.1 Information about the participants

The aim of the first three questions was to acquire information about a person who completed the survey, for a clearer understanding of the context under which the answers were given. Considering the information about the participants, their average age was 41 years, with the

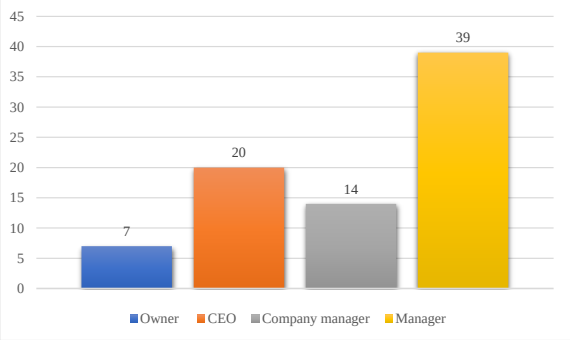



Figure 3: Number of participants for each type of the position inside a company.

oldest one being 60 and the youngest only 26. From 80 participants, 52 were women what represents 65%. The participants' position in the company is shown on Figure 3.

To further back up the validity of the responses it is worth to mention that over 65% of the participants have more than 15 years of working experience, indicating that the participants are sufficiently professional and competent to be the candidates for our survey.

## 4.2 Information about the companies

The companies in Croatia classify, according to their size, into micro (less than 10 employees), small (10 to 50 employees), middle (50 to 250 employees), and big (more than 250 employees).

Figure 4a presents the shares of the company sizes in the number of participants, where it is evident that most of the participated companies are big companies with more than 250 employees. Since larger companies have a more complex organization and usually have proprietary divisions that deal with the welfare of the employees, it is expected that they care and invest more in the education of their employees. Shares of the participated companies according to their output are presented in Figure 4b. As a general trend, most of the companies surveyed are service companies, which are known to utilize ICT in their businesses more than the manufacturing companies.

## 4.3 E-learning practices

Of the companies surveyed, only 13.75% use e-learning for more than ten years, while most of them (71.25%) use it for less than five years. This supports our first hypothesis (H1) that e-learning is fairly new in Croatian businesses. Nevertheless, one should bear in mind that the expansion of e-learning in the Republic of Croatia has only occurred in the last decade. The surveyed companies are evidence of modernizing and digitalizing of employee education. The competency of the employees in the field of ICT is vital for harnessing the benefits of e-learning. This is evident from the fact that, according to participants' opinion, 53.75% of the surveyed companies have more than 50% of employees who have the needed skills to use e-learning platforms, while only 12.5% of the companies have less than 10% of employees with those skills. Consequently, the surveyed companies form a good

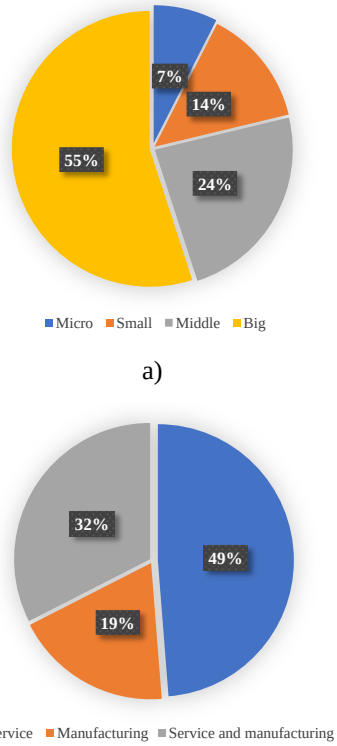



Figure 4: Company classification according to a) size, b) the type of its output.

representative sample for surveying opinions on e-learning. All the participants claim that their company uses the Internet or Intranet for their daily businesses. According to them, the modern business market requires workers with additional skills not obtained via formal education process, where 71% of participants answered that they completely agree, see Figure 5. This is on par with H2, since companies agree that e-learning is necessary and useful.

Investing in new technologies and employee education is important for each company to keep up with the market demands, [19]. A vast majority of the surveyed companies recognize that importance, considering that almost 94% of them invest in educating their employees. The frequency of investing in the education of the company's employees is shown in Figure 6. There are still companies that rarely invest in educating their employees (2.67% of them responded so), which can be due to different reasons.

One of the major reasons could be the fact that they consider their business do not need updating, while the second is probably that their employees are being educated elsewhere.

The frequency of investments in employee education concerning desired outcomes is shown in Figure 7. Most companies focus on investments in upgrading the skills of their employees needed for their current job and, to some extent, to prepare them for the future challenges of their workplaces. Still, most of them periodically invest in career development and education for new working skills.

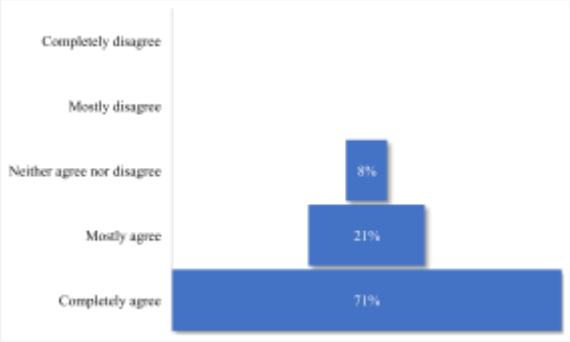



Figure 5: To which extent the participants agree that modern business market requires workers with additional skills not obtained via formal education process.

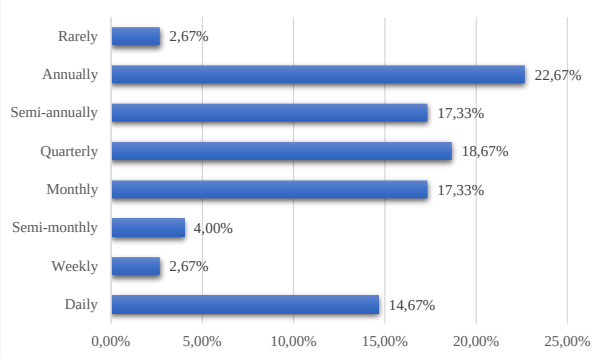



Figure 6: The frequency of investments in employee education.

The types of investments in employee education show that the companies are aware of the future challenges and try to keep up with them as much as their budget allows.

When we correlate the frequency of investing in education with the company size it is evident that smaller companies share similar habits with the big ones (Figure 8.). The only difference is that big companies avoid daily investments in education and prefer to do it annually, whereas smaller companies prefer daily investments. This supports H2 since it is obvious that all companies, regardless of their size, tend to invest in educating their employees. Also, the habits of investing in employee education support H3, since there is a clear awareness that educating employees is useful and that it will eventually pay-off.

Since bigger companies have more complex organizations and specialized divisions that deal with and plan the activities of their employees, they mostly have an elaborate plan for the education of their employees (i.e. there are no spontaneous investments into education). Smaller companies overcome the lack of planning and resources to reorganize quickly and to retrain their employees for the upcoming business demands, therefore having more frequent investments in the education of their workers.

From the survey results, it is evident that the companies recognize the importance of employee education and using the e-learning platforms. The deeper analysis of how the employee education will reflect on

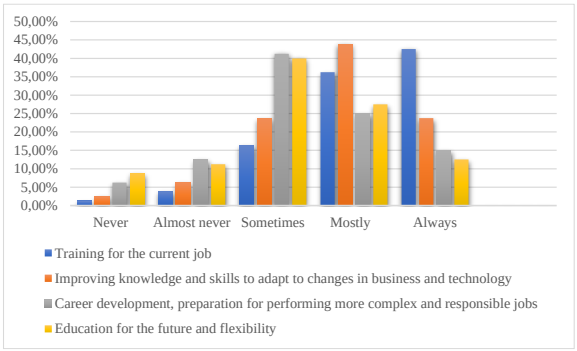



Figure 7: Frequency of investing in employee education according to the desired outputs.

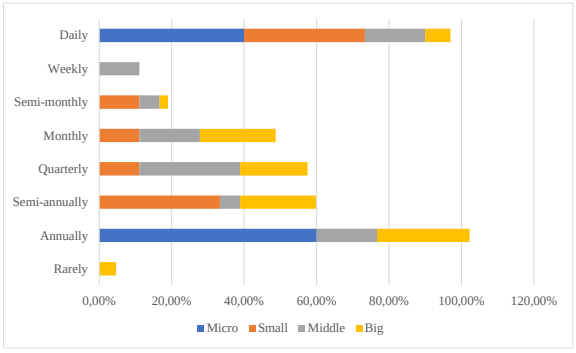



Figure 8: Frequency of investing in employee education according to the company size.

their company, extracted from the participants' opinions, is given in Table 2.

For all the analyzed education outcomes the participants agree that the education will mostly bring benefits. This is especially evident for the claims that education improves problem-solving and that it improves controlling new and unusual situations that may arise in the company. Only one claim had the biggest variety of responses: that the education will improve customer service. This led us to conclude that participants believe more parameters are affecting the customer service, other than education (probably social skills and the like).

Most participants believe that the areas that are least likely to be improved by education are the creativity of their employees and conflict management, but this is only to a lesser extent. This supports H3 since most participants believe that education (served via e-learning platforms) will increase the overall work performance.

## 4.4 The usage of e-learning platforms

Analysis of the frequency of e-learning platform use shows that only a small number of employees use it for daily working tasks or education, while most employees use it only up to two hours monthly. Level of integration of e-learning in daily tasks is still at a low level, thus somewhat supporting H1 (e-learning being a fairly new concept to Croatian companies) and, to some extent, H4 (e-learning appears as an additional load on top of the usual tasks).

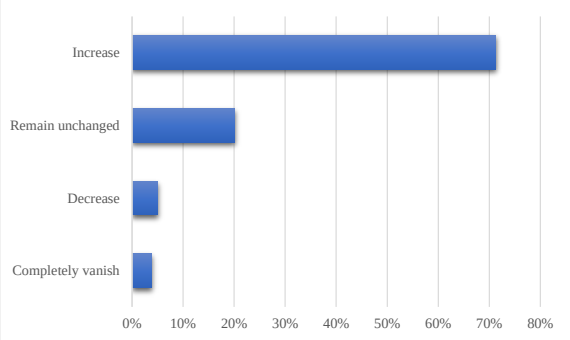



Figure 9: Participants' opinion on e-learning trend.

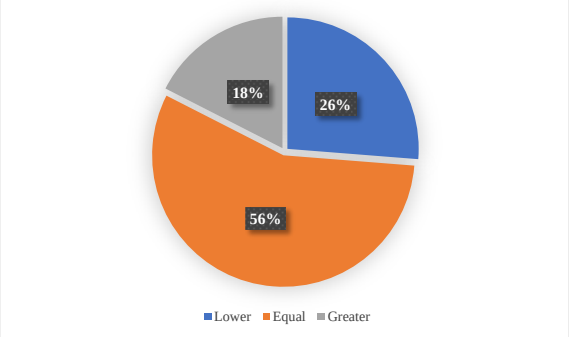



Figure 10: The participants' opinion about the difference in the amount of knowledge acquired using e-learning compared to traditional teaching methods.

Finally, we asked the participants to share their projections on the future of e-learning platforms. Figure 9 gives their responses.

The results indicate that e-learning has a shiny future; it will probably be one of the basic tools used for improving businesses. At the same time, the participants do not think that e-learning is necessarily a great advantage compared to the traditional teaching methods, see Figure 10. The general opinion is that the amount of knowledge acquired is equal regardless of using either e-learning or traditional teaching methods, which was the answer of 56% of the participants. Also, only a minority of participants think that the amount of knowledge acquired via e-learning is greater compared to the traditional teaching methods (18%). Both results support H1, since there is still a dose of distrust towards e-learning platforms, and people still prefer the presence of teachers and other participants.

The results indicate that e-learning has a shiny future; it will probably be one of the basic tools used for improving businesses. At the same time, the participants do not think that e-learning is necessarily a great advantage compared to the traditional teaching methods, see Figure 10. The general opinion is that the amount of knowledge acquired is equal regardless of using either e-learning or traditional teaching methods, which was the answer of 56% of the participants. Also, only a minority of participants think that the amount of knowledge acquired via e-learning is greater compared to the traditional teaching methods (18%). Both results support H1, since there is still a dose of distrust towards e-learning platforms, and people still prefer the presence of teachers and other participants.

Content-wise, e-learning platforms are quite advanced and rich systems. Today's e-learning platforms introduce many different types of content and activities that can be used to produce, share, analyze, verify, and assess the course materials. With that in mind, we asked the participants to analyze which of these types of content and activities they find the most important for acquiring knowledge, and which ones they use the most. Regarding the intensity of use, the results show that games, quizzes, and virtual classrooms are most scarcely used, Figure 11.

Interestingly enough, those very virtual classrooms were the main means of teaching in many academic institutions during the COVID-19 pandemic. This indicates that in the business sector e-learning was used in the same way prior the pandemic and during the pandemic – more oriented towards individual learning. The content

| | Never | Almost never | Sometimes | Mostly | Always |
|---|---|---|---|---|---|
| **Education will improve customer service** | 1.25% | 5.00% | 22.50% | **46.25%** | 25.00% |
| **Education will contribute to change of attitude** | 2.50% | 7.50% | 33.75% | **42.50%** | 13.75% |
| **Education will improve teamwork** | 0.00% | 3.75% | 27.50% | **46.25%** | 22.50% |
| **Education will improve time management** | 1.25% | 5.00% | 25.00% | **47.50%** | 21.25% |
| **Education will improve work safety** | 1.25% | 5.00% | 28.75% | **50.00%** | 15.00% |
| **Education will improve problem solving** | 1.25% | 1.25% | 21.25% | **58.75%** | 17.50% |
| **Education will contribute to better handling of new tasks** | 1.25% | 1.25% | 13.75% | **58.75%** | 25.00% |
| **Education will improve creativity** | 1.25% | 10.00% | 32.50% | **45.00%** | 11.25% |
| **Education will improve product quality** | 1.25% | 3.75% | 26.25% | **51.25%** | 17.50% |
| **Education will improve conflict management** | 0.00% | 10.00% | 28.75% | **48.75%** | 12.50% |
| **Education will drive up to promotion** | 1.25% | 5.00% | 38.75% | **41.25%** | 13.75% |

Table 2: Participants' opinion on how employee education will affect their companies.

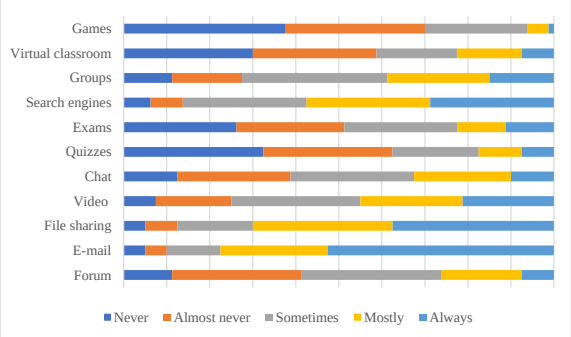



Figure 11: The frequency of using different types of e-learning platform content.

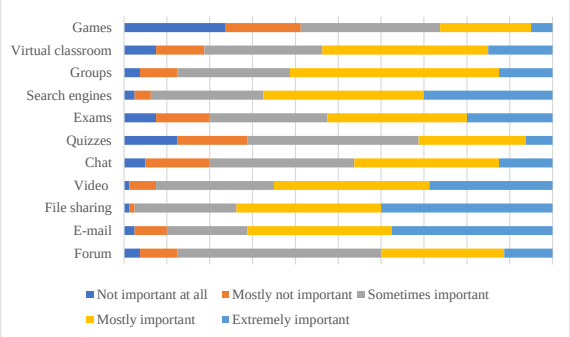



Figure 12: The importance of different types of e-learning platform content.

types that were most often used were emails, followed by file sharing and, to a lesser extent, search engines.

The importance of different types of content and activities within e-learning platforms is given in Figure 12. The participants think that file sharing is always important, following by emails and then search engines. The participants were quite indecisive when determining the importance of forums and quizzes, although these tools are extensively used. But they are used more from the teacher's side than from the participant's side since participants use them more passively (mostly for getting short information).

Comparison of the results shown in Figure 11 and Figure 12 reveals the tight correlation between the importance and intensity of use of different activities within e-learning platforms. This means that the participants believe that the types of content they use the most is the very relevant one for acquiring knowledge, and vice-versa for the content types they use scarcely. The indication is that they are mostly satisfied with the current e-learning processes, but also that they got used to certain type of activities. This somewhat proves H3, since participants believe that e-learning is conducted in a targeted manner.

# 5 Conclusion

In today's dynamic times, one is not limited to the compulsory knowledge acquired by the school system, and on the other hand, employers cannot afford the routine work of their employees. It is extremely important to develop and expand knowledge and to improve and acquire new skills to keep up with new trends and create added value for one's company.

E-learning systems are a fast, reliable, and effective way of teaching which is not limited by space and time provided for teaching. In addition to saving time and reducing training costs, it enforces greater responsibility and self-discipline on participants, it enables self-evaluation of acquired knowledge and progress monitoring, thus ultimately leading to better knowledge transfer. E-learning additionally motivates students compared to the F2F learning by enabling repeated returns to the educational content over a longer period and by choosing the pace of learning adapted to their obligations. E-learning also provides a consistent and standardized approach to each education, thus allowing each participant to gain the same experience regardless of time and place of the course. Via e-learning, an unlimited number of participants can undergo the same training. By utilizing e-learning, companies can offer additional education and the possibility of advancement to their employees, which is a significant motivator today for being loyal to the company that invests in their peoples' education. In crises, such as the current COVID-19, one can see how important it is to keep up with the technological advancement, especially in education. E-learning in companies should be planned to be in line with the strategic goals of the organization. The spread and further application of e-learning in business organizations will depend on meeting their expectations and proving the justification of the investment. This is possible with a planned, thorough, and continuous evaluation of the results of this type of teaching. Hence, it would be significant to continue research in this direction. For the future work, we plan to broaden the analysis by comparing the survey results coming from the different industries, different company sizes and different regions. We aim to obtain even better refinement of the existing appraisal of the e-learning, as well as the possible shortcomings of it. Also, we plan to include more companies in the survey, and further refine the questions themselves. We also plan to complement this research by investigating students' attitude towards e-learning, considering them as the future employees. We will specially target the students coming from technical universities who will eventually be the ones that will further develop e-learning.

Results of the conducted survey show that e-learning in business practice in the Republic of Croatia is still in the early phases of adoption. Despite the companies' opinion that they utilize e-learning in a daily basis, it is a matter of the day when this will become the reality. Furthermore, the results show that they consider investing in employee's education as an important factor for the companies' market competitiveness.

# References

[1] A. S. Tsui, J. L. Pearce, L. W. Porter, and A. M. Tripoli (1997). "Alternative approaches to the employee-organization relationship: does investment in employees pay off?. *Academy of Management Journal*, Academy of management, vol. 40, no. 5, pp. 1089–1121. doi: 10.2307/256928.

[2] D. Altig *et al.* (2020). "Economic uncertainty before and during the COVID-19 pandemic. "*Journal of Public Economics*, Elsevier, vol. 191. doi: 10.1016/j.jpubeco.2020.104274.

[3] A. Kapo, A. Mujkic, L. Turulja, and J. Kovačević (2020). "Continuous e-learning at the workplace: the passport for the future of knowledge". *ITP*, vol. ahead-of-print, no. ahead-of-print. doi: 10.1108/ITP-04-2020-0223.

[4] H. M. Selim (2007). "Critical success factors for e-learning acceptance: Confirmatory factor models". *Computers & Education*, Elsevier, vol. 49, no. 2, pp. 396–413. doi: 10.1016/j.compedu.2005.09.004.

[5] S. Negash, M. Whitman, A. Woszczynski, K. Hoganson, and H. Mattord, Eds (2008). *Handbook of Distance Learning for Real-Time and Asynchronous Information Technology Education.* IGI Global. doi: 10.4018/978-1-59904-964-9

[6] M. Puteh and S. Hussin (2007). "A comparative study of e-learning practices at Malaysian private universities", *1st International Malaysian Educational Technology Convention (2007)*

[7] E. K. Kahiigi, L. Ekenberg, H. Hansson, F. F. T. Danielson, and M. Danielson (2008). "Exploring the e-Learning State of Art". *Electronic Journal of e-Learning*, Academic Conferences and Publishing International vol. 6, no. 2, p. 13.

[8] W. K. Horton (2000), "*Designing Web-based training: how to teach anyone anything anywhere anytime*". New York: Wiley

[9] M. Ćukušić and M. Jandrić (2012), "*E-učenje: koncept i primjena*". Zagreb: Školska knjiga

[10] H. M. W. Rasheed, Y. He, J. Khalid, H. M. U. Khizar, and S. Sharif (2020). "The relationship between e-learning and academic performance of students," *Journal of Public Affairs*, John Wiley & Sons Ltd. doi: 10.1002/pa.2492.

[11] M. M. Škrtić, K. Horvatinčić, and A. Pisarović (2017). *"E-learning from business processes aspect".*

[12] Edward R. Kemery (2000), "*Developing On-Line Collaboration,* IGI Global doi:10.4018/978-1-878289-60-5.ch014

[13] K. Postolov, M. Magdinceva Sopova, and A. Janeska Iliev (2017). "E-learning in the hands of generation Y and Z". *Business excellence*, Faculty of Economic & Business, vol. 11, no. 2, pp. 107–119. doi: 10.22598/pi-be/2017.11.2.107.

[14] D. A. Cook (2007), "Web-based learning: pros, cons and controversies," *Clinical Medicine*, Royal College of Physicians, vol. 7, no. 1, pp. 37–42. doi: 10.7861/clinmedicine.7-1-37.

[15] J. L. Moore, C. Dickson-Deane, and K. Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," *The Internet and Higher Education*, vol. 14, no. 2, pp. 129–135, Mar. 2011, doi: 10.1016/j.iheduc.2010.10.001.

[16] S. Ninoriya, P. M. Chawan, and B. B. Meshram (2011). "CMS, LMS and LCMS For eLearning". *International journal of computer science*, International Journal of Computer Science Issues, vol. 8, no. 2, p. 5.

[17] E. Dahlstrom, D. C. Brooks, and J. Bichsel (2014). "The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives*". EDUCAUSE Annual Conference 2014*, EDUCASE p. 27. doi: 10.13140/RG.2.1.3751.6005.

[18] Y.-M. Cheng (2014). "Roles of interactivity and usage experience in e-learning acceptance: a longitudinal study". *International Journal of Web Information Systems*, Emerald Publishing Limited vol. 10, no. 1, pp. 2–23. doi: 10.1108/IJWIS-05-2013-0015.

[19] M. Di Ubaldo and I. Siedschlag (2020). "Investment in Knowledge-Based Capital and Productivity: Firm-Level Evidence from a Small Open Economy". *Review of Income and Wealth*, International Association for Research in Income and Wealth doi: 10.1111/roiw.12464.

# A Semi-Supervised Approach to Monocular Depth Estimation, Depth Refinement, and Semantic Segmentation of Driving Scenes using a Siamese Triple Decoder Architecture

John Paul T. Yusiong[1,2] and Prospero C. Naval, Jr.[1]
[1]Computer Vision and Machine Intelligence Group, Department of Computer Science
College of Engineering, University of the Philippines, Diliman, Quezon City, Philippines
[2]Division of Natural Sciences and Mathematics
University of the Philippines Visayas Tacloban College, Tacloban City, Leyte, Philippines
E-mail: jtyusiong@up.edu.ph; pcnaval@up.edu.ph

**Keywords:** Siamese triple decoder architecture, depth estimation and refinement, semantic segmentation, semi-supervised learning methods

**Received:** November 30, 2019

*Depth estimation and semantic segmentation are two fundamental tasks in scene understanding. These two tasks are usually solved separately, although they have complementary properties and are highly correlated. Jointly solving these two tasks is very beneficial for real-world applications that require both geometric and semantic information. Within this context, the paper presents a unified learning framework for generating a refined depth estimation map and semantic segmentation map given a single image. Specifically, this paper proposes a novel architecture called JDSNet. JDSNet is a Siamese triple decoder architecture that can simultaneously perform depth estimation, depth refinement, and semantic labeling of a scene from an image by exploiting the interaction between depth and semantic information. A semi-supervised method is used to train JDSNet to learn features for both tasks where geometry-based image reconstruction methods are employed instead of ground-truth depth labels for the depth estimation task while ground-truth semantic labels are required for the semantic segmentation task. This work uses the KITTI driving dataset to evaluate the effectiveness of the proposed approach. The experimental results show that the proposed approach achieves excellent performance on both tasks, and these indicate that the model can effectively utilize both geometric and semantic information.*

*Povzetek: V članku je predstavljena izvirna metoda delno nadzorovanega učenja za raznovrstne vizualne naloge.*

## 1 Introduction

Scene understanding is crucial for autonomous driving systems since it provides a mechanism to understand the scene layout of the environment [1, 2]. Scene understanding involves depth estimation and semantic segmentation, which facilitates the understanding of the geometric and semantic properties of a scene, respectively. Depth estimation and semantic segmentation address different areas in scene understanding but have complementary properties and are highly correlated.

For semantic segmentation, depth values help improve semantic understanding by enabling the model to generate more accurate object boundaries or differentiate objects having a similar appearance since these values encode structural information of the scene. On the other hand, for depth estimation, the semantic labels provide valuable prior knowledge to depict the geometric relationships between pixels of different classes and generate better scene layout [3, 4, 5, 6]. Thus, these two fundamental tasks in computer vision can be dealt with in an integrated manner under a unified framework that optimizes multiple objectives to improve computational efficiency and performance for both tasks from single RGB images. However, addressing depth estimation and semantic segmentation simultaneously where the two tasks can benefit from each other is non-trivial and is one of the most challenging tasks in computer vision given the peculiarities of each task and the limited information that can be obtained from monocular images.

Previous works jointly model these two tasks using traditional hand-crafted features and RGB-D images [7, 8]. However, the hand-crafted feature extraction process is quite tedious, and it generally fails to help achieve high accuracies while RGB-D image acquisition is a costly endeavor. To overcome the aforementioned issues, researchers employ a unified framework based on deep learning that enables these two tasks to enhance each other using single RGB images only, and this approach led to a significant breakthrough for both tasks [4, 5, 6, 9, 10, 11, 12]. Since these unified frameworks are based on the fully-supervised learning method, they require vast quantities

of training images with per-pixel ground-truth semantic labels and depth measurements, and obtaining these ground-truths is non-trivial, costly, and labor-intensive. An alternative approach, as proposed by Ramirez *et al.* [13], is to integrate depth estimation and semantic segmentation into a unified framework using the semi-supervised learning method. The semi-supervised learning framework requires ground-truth semantic labels to provide supervisory signals for the semantic segmentation task, while for the depth estimation task, it employs geometry-based image reconstruction methods that utilize secondary information based on the underlying theory of epipolar constraints instead of requiring ground-truth depth measurements during training. In other words, addressing the problem of scene understanding assumes that both stereo image pairs and semantic information are available during training since this framework exploits the relationship between the geometric and semantic properties of a scene by performing semantic segmentation in a supervised manner and casting the depth estimation task as an image reconstruction problem in an unsupervised manner.

This paper presents another attempt towards addressing the joint inference problem involving depth estimation and semantic segmentation from a single image by proposing to train a novel architecture using a unified learning framework based on a semi-supervised technique. This paper introduces a novel Siamese triple decoder architecture with a disparity refinement module and a segmentation fusion module. The triple decoder architecture consists of one shared encoder and three parallel decoders. The disparity refinement module handles visual artifacts and blurred boundaries to generate better depth maps with no border artifacts around the image boundary while the segmentation fusion module generates the semantic segmentation map. In contrast, previous works apply a non-trainable post-processing heuristic during testing to refine the depth estimation outputs of the trained model [13, 14]. Essentially, the proposed method enables the model to simultaneously perform depth estimation, depth refinement, and semantic labeling of a scene from an image by exploiting the interaction between depth and semantic information in an end-to-end manner.

The main contributions of this work are the following:

1. It introduces a novel Siamese triple decoder architecture with a disparity refinement module and a segmentation fusion module, referred to as JDSNet, for depth estimation, depth refinement, and semantic segmentation.

2. It presents a unified framework for joint depth estimation with depth refinement and semantic segmentation from a single image based on a semi-supervised technique and trains JDSNet to simultaneously perform depth estimation, depth refinement, and semantic segmentation in an end-to-end manner using rectified stereo image pairs with ground-truth semantic labels as training data.

3. It describes a training loss function that optimizes these two tasks concurrently.

4. It demonstrates that the proposed method is capable of simultaneously addressing these two tasks that are mutually beneficial to both tasks. The experimental results prove that jointly solving these two tasks improves the performance of both tasks on various evaluation metrics.

The remainder of the paper is arranged as follows. Section 2 introduces the related works. Section 3 describes the proposed semi-supervised learning framework for simultaneous monocular depth estimation, depth refinement, and semantic segmentation. Section 4 discusses the experimental results using a standard benchmark dataset. Lastly, Section 5 concludes the paper.

## 2 Related work

This section focuses on the previous works that dealt with joint depth estimation and semantic segmentation where researchers attempted to develop better-suited models using different methods, such as traditional hand-crafted feature extraction techniques and deep learning-based techniques.

The earliest works [7, 8] show the feasibility of jointly modeling depth estimation and semantic segmentation from a single RGB image using the supervised learning method. However, they employ traditional hand-crafted features for these two tasks. The work of Ladicky *et al.* [7] is considered to be the first to jointly perform monocular depth estimation and semantic segmentation. Using properties of perspective geometry, they proposed an unbiased semantic depth classifier and considered both the loss from semantic and depth labels when training the classifier. They obtained results that outperformed previous state-of-the-art traditional methods in both the monocular depth and semantic segmentation domain. But, their model can only generate coarse depth and semantic segmentation maps because the predictions are based on local regions with hand-crafted features. Similarly, Liu *et al.* [8] carried out these two tasks in a sequential manner where they first performed semantic segmentation and then used the predicted semantic labels to improve the depth estimation accuracy. Specifically, they used Markov Random Field (MRF) models for depth estimation, where a multi-class image labeling MRF predicts the semantic class for every pixel in the image and uses the predicted semantic labels as priors to estimate depth for each class. By incorporating semantic features, they achieved excellent results with a simpler model that can take into account the appearance and geometry constraints.

Other researchers [6, 12, 13] use deep learning techniques for joint monocular depth estimation and semantic segmentation from a single image to improve the performance of each task. These works [6, 12] performed depth estimation and semantic labeling using the super-

vised learning method while Ramirez *et al.* [13] used the semi-supervised learning method.

Wang *et al.* [6] and Mousavian *et al.* [12] used deep network architecture to simultaneously perform depth estimation and semantic segmentation and used a Conditional Random Field (CRF) to combine the depth and semantic information. Specifically, Wang *et al.* [6] proposed a two-layer Hierarchical Conditional Random Field (HCRF), which employs two convolutional neural networks (CNNs) to extract local and global features and then these features are enhanced using CRF. Their proposed approach enabled them to obtain promising results in both the monocular depth and semantic segmentation domain. On the other hand, Mousavian *et al.* [12] introduced a multi-scale CNN to perform depth estimation and semantic segmentation and combined them using a CRF. As shown in their work, the proposed model achieved comparable results on monocular depth estimation but outperformed the state-of-the-art methods on semantic segmentation. A more recent work by Ramirez *et al.* [13] proposed to solve the joint inference problem using a semi-supervised learning method where they employed a deep network architecture that can be jointly optimized for depth estimation and semantic segmentation where ground-truth semantic labels are required for the semantic segmentation task while geometry-based image reconstruction methods are employed instead of ground-truth depth labels for the depth estimation task. However, the experimental results reveal that their model, which was jointly trained for depth prediction and semantic segmentation, only improved the depth estimation accuracy. Their model failed to obtain better results for semantic segmentation.

This work addresses past design issues to obtain significant improvements when simultaneously performing depth estimation and semantic segmentation using rectified stereo image pairs with ground-truth semantic labels as training data. Specifically, to produce better depth estimates and semantic labeling, the proposed method involves changing the essential building blocks of the network architecture and introducing a disparity refinement module and a segmentation fusion module to generate better quality depth maps and semantic segmentation maps.

# 3 Proposed method

This section describes the proposed method for simultaneous depth estimation, depth refinement, and semantic segmentation in a semi-supervised manner using rectified stereo image pairs $(I_L,\ I_R)$ with ground-truth semantic labels $seg^{gt}$ as training data. Since the training data does not have ground-truth depth labels, the right images $I_R$ together with the predicted disparities $D_{L1}$ are used to obtain supervisory signals for the depth estimation task based on the underlying theory of epipolar constraints during training. In short, the supervisory signal is generated by warping one view of a stereo pair into the other view using the predicted disparity maps. Figure 1 presents the semi-supervised framework for joint monocular depth estimation and semantic segmentation using JDSNet. JDSNet is the proposed Siamese triple decoder architecture with a disparity refinement module and a segmentation fusion module.

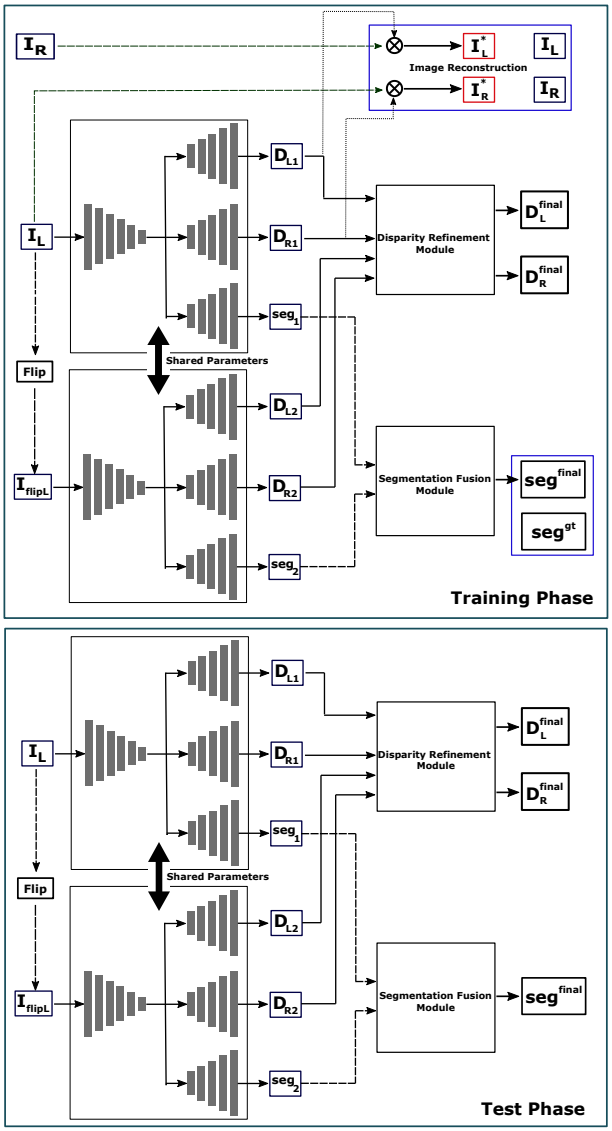



Figure 1: A semi-supervised framework for joint monocular depth estimation and semantic segmentation using JDSNet, the proposed Siamese triple decoder architecture.

## 3.1 Network architecture

The semi-supervised framework uses a Siamese architecture with the triple decoder network as the autoencoder. A Siamese architecture consists of two symmetrical structures and accepts two distinct images as inputs. An important feature of a Siamese architecture is that it uses two copies of the same network, and these two networks share weight parameters to process the two different inputs and generate two outputs. The original purpose of using a

Siamese architecture is for learning similarity representations, that is, to predict whether the two inputs are similar or not [15, 16]. However, in this study, the two outputs of the Siamese network are combined to produce the refined disparity maps through the disparity refinement module and a segmentation map through the segmentation fusion module.

JDSNet consists of two triple decoder networks that share weight parameters. It also has a disparity refinement module that enables the network to more effectively handle the visual artifacts and blurred boundaries while learning depth estimation. The disparity refinement module is the trainable version of the post-processing heuristic introduced by Godard *et al.* [14]. This module combines and refines the two pairs of depth maps. The segmentation fusion module combines the outputs from the two semantic segmentation decoders.

The Siamese triple decoder network receives the original left images $I_L$ and the horizontally flipped version of the left input images $I_{flipL}$ as inputs. With these images, the network is trained to predict depth maps, refine the predicted depth maps, and generate semantic segmentation maps.

The horizontally flipped version of the left input images $I_{flipL}$ is necessary because in reconstructing the left images from the right images using the predicted disparities, there are pixels in the left images that are not present in the right images. Hence, no depth values can be predicted for these missing pixels. To overcome this limitation, the horizontally flipped version of the left input images $I_{flipL}$ enables the network to predict the depth values of the occluded pixels, and by using the disparity refinement module, the predicted disparities from both inputs are combined to generate a refined disparity map.

A triple decoder network has a shared encoder and three parallel decoders that can be trained for depth estimation and semantic segmentation. The shared encoder is based on the encoder section of the AsiANet network architecture [17]. The encoded feature vectors are forwarded to the three parallel decoders: two depth decoders and one semantic segmentation decoder. The first depth decoder predicts the left disparity map and is constructed similar to the decoder section of AsiANet [17], while the second depth decoder that predicts the right disparity map and the semantic segmentation decoder are based on the ResNet50 decoders described in [13]. However, the last encoder block is modified due to hardware limitations where the number of output channels is reduced from 2048 to 1024. Also, unlike the previous works [13, 17], where a depth decoder generates two disparity maps when using rectified stereo image pairs as training data, each depth decoder in the proposed network generates a single disparity map.

The Siamese triple decoder network generates a pair of refined disparity maps ($D_L^{final}$, $D_R^{final}$) at four different scales and a semantic segmentation map $seg^{final}$ at full resolution only from the left image $I_L$. However, only the full resolution of the refined left disparity map $D_L^{final}$ and semantic segmentation map are useful at test time.

#### 3.1.1 Disparity refinement module

The disparity refinement module is based on the post-processing heuristic introduced by Godard *et al.* [14]. It is incorporated as a trainable component of the proposed Siamese triple decoder network rather than having a refinement step at test time since it decouples the refined disparity maps from the training. This design choice enables the network to simultaneously learn depth estimation and refine the predicted depth map in an end-to-end manner.

Essentially, the disparity refinement module performs three operations: horizontal flip operation, pixel-wise mean operation, and disparity ramps removal operation. The horizontal flip operation is performed on the disparity maps ($D_{L2}$, $D_{R2}$) to generate ($D_{flipL}$, $D_{flipR}$). Afterwards, the pixel-wise mean operation and the disparity ramps removal operation are performed on ($D_{L1}$, $D_{flipL}$) and ($D_{R1}$, $D_{flipR}$), respectively, to produce the refined disparity maps ($D_L^{final}$, $D_R^{final}$).

#### 3.1.2 Segmentation fusion module

The segmentation fusion module performs a horizontal flip operation on $seg_2$ to obtain $seg_{flip}$. It then adds the two layers $seg_1$ and $seg_{flip}$ and forwards it to the softmax layer to output the probabilistic scores for each class and generate a semantic segmentation map $seg^{final}$.

### 3.2 Loss function

Training the proposed network relies on a loss function that can be expressed as a weighted sum of two losses, as defined in equation (1); a depth loss and a semantic segmentation loss, and the term is given by

$$L_{Total} = \alpha_{depth} L_{depth} + \alpha_{seg} L_{seg}, \tag{1}$$

where $L_{depth}$ is the depth loss term, $L_{seg}$ is the semantic segmentation loss term, and $\alpha_{depth}$, $\alpha_{seg}$ are the loss weightings for each term.

#### 3.2.1 Depth loss term

As defined in equation (2), $L_{depth}$ is the sum of the depth losses at four different scales where $L_s$ is the depth loss at each scale. $L_s$ is a combination of three terms - appearance dissimilarity, disparity smoothness, and left-right consistency. This term is given by

$$L_{depth} = \sum_{s=1}^{4} L_s, \tag{2}$$

$$L_s = \alpha_{app} L_{app} + \alpha_{sm} L_{sm} + \alpha_{lr} L_{lr}, \tag{3}$$

$$L_{app} = L_{app}^{left} + L_{app}^{right}, \tag{4}$$

$$L_{sm} = L_{sm}^{left} + L_{sm}^{right}, \tag{5}$$

$$L_{lr} = L_{lr}^{left} + L_{lr}^{right}, \tag{6}$$

where $L_{app}$ is the appearance dissimilarity term, $L_{sm}$ is the edge-aware disparity smoothness term, $L_{lr}$ is the left-right consistency term, and $\alpha_{app}, \alpha_{sm}, \alpha_{lr}$ are the loss weightings for each term. The depth loss term takes into account the left and right images where each component is in terms of the left images ($L_{app}^{left}, L_{sm}^{left}, L_{lr}^{left}$) and right images ($L_{app}^{right}, L_{sm}^{right}, L_{lr}^{right}$). However, this section provides details for the left components $L^{left}$ only since the right components $L^{right}$ are defined symmetrically.

The appearance dissimilarity term, as defined in (7), is a linear combination of the single-scale structural similarity (SSIM) term [18] and the $L_1$ photometric term. This term measures the quality of the synthesized target image by minimizing the pixel-level dissimilarity between the target image $I$ and the synthesized target image $I^*$. This term is also widely used in previous studies [13, 14, 17] and it is given by

$$L_{app}^{left} = \frac{1}{N}\sum_{x,y}\omega\frac{1-SSIM\left(I_L\left(x,y\right),I_L^*\left(x,y\right)\right)}{2} + \left(1-\omega\right)\left\|I_L\left(x,y\right)-I_L^*\left(x,y\right)\right\| \tag{7}$$

with a $3 \times 3$ box filter for the SSIM term and $\omega$ is set to $0.85$ similar to [13, 14, 17]. The synthesized target left image $I_L^*$ is obtained using a sampler from the spatial transformer network [19] that performs the bilinear interpolation. The sampler reconstructs the target left image $I_L^*$ using the right image $I_R$ and the predicted left disparity map $D_{L1}$.

The edge-aware disparity smoothness term, as defined in (8), regularizes the predicted disparities in spatially similar areas to ensure that the predicted disparities are locally smooth but can be sharp at the edges. This term is given by

$$\begin{aligned}{}_{sm}^{left} =& \frac{1}{N}\sum_{x,y}((|\partial_x D_{L2}(x,y)|e^{-|\partial_x I_{flipL}(x,y)|} \\ &+ |\partial_y D_{L2}(x,y)|e^{-|\partial_y I_{flipL}(x,y)|}) \\ &+ (|\partial_x D_L^{final}(x,y)|e^{-|\partial_x I_L(x,y)|} \\ &+ |\partial_y D_L^{final}(x,y)|e^{-|\partial_y I_L(x,y)|})),\end{aligned} \tag{8}$$

where $D_L^{final}$ is the refined left disparity map, $D_{L2}$ is the second predicted left disparity map, and $I_{flipL}$ is the horizontally flipped version of the left image $I_L$.

As described in [13, 14, 17], the left-right consistency term enforces consistency between the left and right disparities as defined in (9). This term is given by

$$\begin{aligned}L_{lr}^{left} =& \frac{1}{N}\sum_{x,y}|D_L^{final}(x,y) \\ &- (D_{R1}(x - D_{L1}(x,y),y))|,\end{aligned} \tag{9}$$

where $D_L^{final}$ is the refined left disparity map, $D_{R1}$ is the first predicted right disparity map, and $D_{L1}$ is the first predicted left disparity map.

### 3.2.2 Semantic segmentation loss term

The semantic segmentation loss term, as defined in equation (10), is the standard cross-entropy loss between the predicted pixel-wise semantic labels $seg^{final}$ and ground-truth pixel-wise semantic labels $seg^{gt}$. The semantic segmentation loss is computed using the left images only since these images have the corresponding ground-truth semantic labels at full image resolution. This term is given by

$$L_{seg} = -\sum_{i=1}^{N} P(seg_i^{gt}|seg_i^{final}), \tag{10}$$

where $seg_i^{final}$ is the pixel-wise prediction for image $I_i$, $seg_i^{gt}$ is the ground-truth semantic labels for image $I_i$, $P(y|x) = \sum_j p(y_j|x_j)$, and $p(y_j|x_j)$ is the probability of the ground-truth semantic label $y_j$ at pixel $j$.

## 3.3 Datasets and evaluation metrics

Although the Cityscapes dataset [20] and KITTI dataset [21] contain a large number of training samples, the proposed semi-supervised learning framework for simultaneous depth estimation, depth refinement, and semantic segmentation requires rectified stereo image pairs with pixel-wise ground-truth semantic labels at training time. Hence, a subset of the Cityscapes dataset, which contains $2,975$ finely annotated images and the KITTI dataset consisting of 200 images with pixel-wise semantic ground-truth labels are used in this work.

Ramirez *et al.* [13] introduced a train/test split from the 200 images of the KITTI dataset for joint depth estimation and semantic segmentation. This dataset was split into 160 samples for the train set and 40 samples for the test set. The test set of 40 samples was used to quantitatively evaluate the proposed method given the distance range of 0-80 meters.

The standard evaluation metrics are used to evaluate the trained models quantitatively. The standard evaluation metrics for depth estimation measure the average errors, where lower values are better and accuracy scores where higher values are preferred [14, 22]. The six standard metrics for depth estimation are absolute relative difference (ARD), square relative difference (SRD), linear root mean square error (RMSE-linear), log root mean square error (RMSE-log), and the percentage of pixels (accuracy score) with thresholds ($t$) of $1.25$, $1.25^2$, and $1.25^3$. These metrics are defined in Eq. (11) to Eq.(15).

$$ARD = \frac{1}{N}\sum\frac{|d_i^p - d_i^g|}{d_i^g} \tag{11}$$

$$SRD = \frac{1}{N}\sum\frac{||d_i^p - d_i^g||^2}{d_i^g} \tag{12}$$

$$RMSE-linear = \sqrt{\frac{1}{N}\sum||d_i^p - d_i^g||^2} \tag{13}$$

$$RMSE-log = \sqrt{\frac{1}{N}\sum||log(d_i^p) - log(d_i^g)||^2} \tag{14}$$

$$\delta < t = percent\ of\ d_i^p\ s.t.\ max\{\frac{d_i^p}{d_i^g}, \frac{d_i^g}{d_i^p}\} \tag{15}$$

$d^g$ and $d^p$ represent the ground-truth and estimated depth, respectively. $N$ represents the number of pixels with valid depth value in the ground truth depth map.

On the other hand, the mean intersection over union (mIoU) is used to evaluate the semantic predictions of the model. It is the standard metric for segmentation tasks. The IoU measures the similarity between the intersection and union of the predicted pixel-wise semantic labels $seg^{final}$ and ground-truth pixel-wise semantic labels $seg^{gt}$, and is calculated on a per-class basis and then averaged, as defined in (16). It is the ratio between the number of true positives (intersection) over the sum of true positives (TP), false positives (FP) and false negatives (FN) (union). This is given by

$$mIoU = \frac{1}{n_{cl}} \sum_{c} \frac{TP_c}{TP_c + FP_c + FN_c}, \tag{16}$$

where $n_{cl}$ is the total number of classes and $c \in 0...n_{cl} - 1$.

Moreover, the pixel accuracy, as defined in (17), was also used to evaluate the performance of the model on the semantic segmentation task since the previous work [13] used this metric. This term is given by

$$pixel\ accuracy = \frac{1}{N} \sum_{c} TP_c, \tag{17}$$

where $TP$ represents the true positives or correctly predicted pixels and $N$ is the total number of annotated pixels.

# 4 Experiments

Tensorflow [23] was used to implement JDSNet. Training the network was performed on a single Nvidia GTX 1080 Ti GPU with 11 GB of memory. The training protocol was similar to [13, 14, 17] where the Adam optimizer [24] with $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-8}$ optimized the model for 50 epochs using the Cityscapes dataset and fine-tuned the model for another 50 epochs using the KITTI 2015 dataset by minimizing the training loss. For training and fine-tuning the model, the learning rate was initially set to $\lambda = 10^{-4}$ for the first 30 epochs and was reduced by half every 10 epoch until the process was completed. Moreover, the training phase involved using the same train/test split introduced in [13], resizing the input images to 256 by 512, using a batch size of 2, and performing data augmentation on the input images. The hyper-parameters have the following values: $\alpha_{depth} = 1.0, \alpha_{seg} = 0.1, \alpha_{app} = 1.0, \alpha_{lr} = 1.0$, and $\alpha_{sm} = 0.1/2^s$, where $s$ is the down-sampling factor ranging from 0 to 3.

## 4.1 Results and discussion

This section discusses the results of the experiments conducted to evaluate the proposed method that simultaneously performs depth estimation, depth refinement, and semantic segmentation. The model was evaluated using the publicly available KITTI 2015 dataset [21] based on the test split introduced in [13]. Each test image has a corresponding ground-truth depth and semantic ground-truth labels.

The experiments involved training three different models:

1. Depth only model: $L_{Total} = L_{depth}$,
2. Semantic only model: $L_{Total} = L_{seg}$, and
3. Depth+Semantic model: Equation (1), which is the proposed training loss function.

In the depth only model, the semantic features are not considered during training. Hence, the model can only predict depth maps. In this setup, the two segmentation decoders and the segmentation fusion module are disabled. On the other hand, in the semantic only model, the depth features are not considered during training. Thus, the model can only generate semantic segmentation maps since the four depth decoders and the disparity refinement module are disabled. The main experiment involved training a depth+semantic model using the proposed method where both the semantic and depth features are considered during training.

Table 1 and Table 2 report the quantitative results. The experiment results were compared with the previous methods by directly using the results reported in [13]. These results reveal the effectiveness of the proposed method, which involved training the model to perform depth estimation, depth refinement, and semantic segmentation simultaneously.

As shown in Table 1, JDSNet is a better-suited model for depth estimation even when trained without any semantic information since it outperformed all previous models that were trained using both depth and semantic information based on the different evaluation metrics. The results also show further improvement when semantic information was considered in training JDSNet. Moreover, lower errors indicate that there are few outliers in the predicted depth maps.

A similar trend can be observed in Table 2, where JDSNet outperformed the previous models in terms of the semantic segmentation task when trained using both depth and semantic information. These results indicate that a good network design can significantly improve the performance of a model for both tasks, and including additional features during training can lead to better results. Specifically, simultaneously training the network for both tasks is more beneficial as the model can achieve better results than training a separate network for each task.

Although the results showed that the JDSNet-trained model using both depth and semantic information achieved high pixel accuracy rating, further validation was necessary since the pixel accuracy metric can be biased by imbalanced datasets. To overcome this limitation, the Jaccard index, also referred to as intersection-over-union, was employed. This evaluation metric takes into consideration

| Method | Error Metric (Lower Is Better) | | | | Accuracy Metric (Higher Is Better) | | |
|---|---|---|---|---|---|---|---|
| | ***ARD*** | ***SRD*** | ***RMSE (linear)*** | ***RMSE (log)*** | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Zhou *et al.* [25] | 0.286 | 7.009 | 8.377 | 0.320 | 0.691 | 0.854 | 0.929 |
| Mahjourian *et al.* [26] | 0.235 | 2.857 | 7.202 | 0.302 | 0.710 | 0.866 | 0.935 |
| GeoNet [27] | 0.236 | 3.345 | 7.132 | 0.279 | 0.714 | 0.903 | 0.950 |
| Godard *et al.* [14] | 0.159 | 2.411 | 6.822 | 0.239 | 0.830 | 0.930 | 0.967 |
| Ramirez *et al.* (ResNet50) [13] | 0.143 | 2.161 | 6.526 | 0.222 | 0.850 | 0.939 | 0.972 |
| Ramirez *et al.* (ResNet50+pp) [13] | 0.136 | 1.872 | 6.127 | 0.210 | 0.854 | 0.945 | 0.976 |
| Ours (JDSNet): Depth only | 0.117 | 1.436 | 5.526 | 0.187 | 0.877 | 0.956 | 0.981 |
| **Ours (JDSNet): Depth+Semantic** | **0.108** | **1.221** | **5.309** | **0.178** | **0.883** | **0.959** | **0.985** |

Table 1: Monocular depth estimation results using the KITTI test split introduced in [13]. The **bold** values indicate the best results.

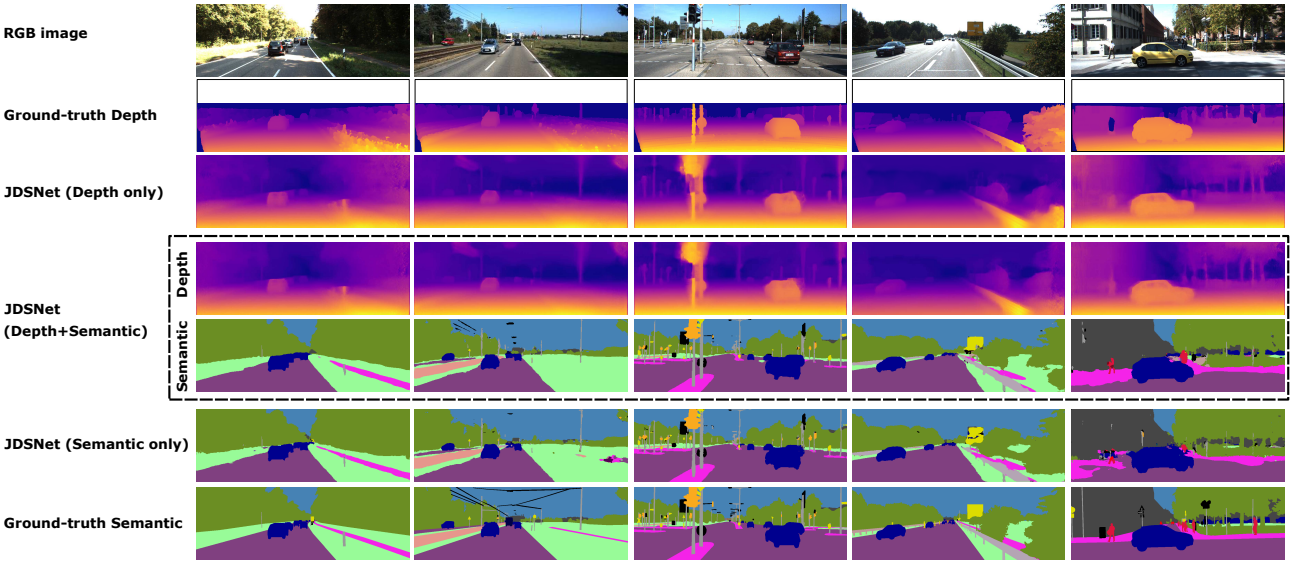



Figure 2: Qualitative results using the KITTI test split introduced in [13]. The ground-truth depth maps are interpolated for visualization purposes only. Best viewed in color.

| Method | *PA* |
|---|---|
| Ramirez *et al.* (ResNet50) [13]: Semantic only | 88.18% |
| Ramirez *et al.* (ResNet50) [13]: Depth+Semantic | 88.19% |
| Ours (JDSNet): Semantic only | 88.40% |
| **Ours (JDSNet): Depth+Semantic** | **89.57%** |

Table 2: Semantic segmentation results using the KITTI test split introduced in [13]. PA means pixel accuracy. The **bold** values indicate the best results.

both the false positives and false negatives. Table 3 confirms that by incorporating depth information JDSNet performed better in the semantic segmentation task. For instance, when using both depth and semantic information, JDSNet was very effective in differentiating ambiguous pairs of classes, such as wall versus fence, sidewalk versus road, and motorcycle versus bicycle. It also achieved better results in terms of recognizing a person and segmenting distant objects and thin structures such as poles, traffic lights, and traffic signs.

The qualitative results, as shown in Figure 2, reveal that the proposed method generated depth maps that captures and preserves the general scene layout where thin structures are perceivable. Also, the disparity refinement module achieved a similar result to the post-processing heuristic that is performed during testing where the refined depth maps have no border artifacts on the image boundary. In addition, the results show that JDSNet can effectively perform semantic segmentation, as evidenced by its ability to capture the geometrical characteristics of the objects in the scene. For example, JDSNet was able to segment the traffic light in the third image even if it has a thin structure and an irregular shape.

# 5 Conclusion

This work has introduced a semi-supervised learning framework that simultaneously performs depth estimation, depth refinement, and semantic segmentation using rectified stereo image pairs with ground-truth semantic labels during training. The proposed architecture, referred to as

| Method | Road | Sidewalk | Building | Wall | Fence | Pole | Traffic light | Traffic sign | Vegetation | Terrain |
|---|---|---|---|---|---|---|---|---|---|---|
| JDSNet: Semantic only | 90.77 | 47.13 | 72.86 | 16.71 | 11.70 | 31.73 | 13.96 | 19.79 | 86.38 | **74.21** |
| JDSNet: Depth+Semantic model | **91.43** | **52.98** | **78.81** | **34.82** | **28.93** | **36.72** | **14.05** | **26.45** | **86.67** | 74.08 |

| Method | Sky | Person | Rider | Car | Truck | Bus | Train | Motorcycle | Bicycle | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|
| JDSNet: Semantic only | 92.83 | 7.68 | **3.09** | 81.82 | 5.18 | **0.00** | **5.04** | 0.31 | 13.88 | 35.53% |
| JDSNet: Depth+Semantic model | **93.37** | **16.87** | 0.74 | **85.65** | **5.23** | **0.00** | 1.51 | **1.53** | **16.85** | **39.30%** |

Table 3: Semantic segmentation results using the KITTI test split introduced in [13]. mIoU means mean intersection over union. The **bold** values indicate the best results.

JDSNet, is a Siamese triple decoder network architecture with a disparity refinement module and a segmentation fusion module that is capable of improving on the performance of both tasks by sharing the underlying features representations and utilizing both geometric and semantic information. Experiment results show that the proposed method achieved promising results on both depth estimation and semantic segmentation and outperformed previous methods.

# References


[1] L. Chen, Z. Yang, J. Ma, and Z. Luo (2018) Driving Scene Perception Network: Real-time Joint Detection, Depth Estimation and Semantic Segmentation, *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, IEEE, pp. 1283–1291. https://doi.org/10.1109/WACV.2018.00145

[2] G. Giannone and B. Chidlovskii (2019) Learning Common Representation from RGB and Depth Images, *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, pp. 408–415. https://doi.org/10.1109/cvprw.2019.00054

[3] R. Cipolla, Y. Gal and A. Kendall (2018) Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 7482–7491. https://doi.org/10.1109/CVPR.2018.00781

[4] J. Liu, Y. Wang, Y. Li, J. Fu, J. Li, and H. Lu (2018) Collaborative Deconvolutional Neural Networks for Joint Depth Estimation and Semantic Segmentation, *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, vol. 29, no. 11, pp. 5655–5666. https://doi.org/10.1109/TNNLS.2017.2787781

[5] D. Sanchez-Escobedo, X. Lin, J. R. Casas, and M. Pardas (2018) Hybridnet for Depth Estimation and Semantic Segmentation, *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 1563–1567. https://doi.org/10.1109/ICASSP.2018.8462433

[6] Peng Wang, Xiaohui Shen, Zhe Lin, S. Cohen, B. Price, and A. Yuille (2015) Towards unified depth and semantic prediction from a single image, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 2800–2809. https://doi.org/10.1109/CVPR.2015.7298897

[7] L. Ladicky, J. Shi, and M. Pollefeys (2014) Pulling things out of perspective, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 89–96. https://doi.org/10.1109/CVPR.2014.19

[8] B. Liu, S. Gould, and D. Koller (2010) Single image depth estimation from predicted semantic labels, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1253–1260. https://doi.org/10.1109/CVPR.2010.5539823

[9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers (2016) Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, *Proceedings of the Asian Conference on Computer Vision*, Springer, pp. 213–228. https://doi.org/10.1007/978-3-319-54181-5_14

[10] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother (2017) Analyzing modular CNN architectures for joint depth prediction and semantic segmentation, *Proceedings of the 2017 International Conference on Robotics and Automation*, IEEE, pp. 4620–4627. https://doi.org/10.1109/ICRA.2017.7989537

[11] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen and I. Reid (2019) Real-Time Joint Semantic Segmentation and Depth Estimation Using Asymmetric Annotations, *Proceedings of the 2019 International Conference on Robotics and Automation*, IEEE, pp. 7101–7107. https://doi.org/10.1109/ICRA.2019.8794220

[12] A. Mousavian, H. Pirsiavash, and J. Košecká (2019) Joint Semantic Segmentation and Depth Estimation with Deep Convolutional Networks, *Proceedings of the 2016 Fourth International Conference on 3D Vision*, IEEE, pp. 611–619. https://doi.org/10.1109/3DV.2016.69

[13] P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano (2018) Geometry meets semantic for semi-supervised monocular depth estimation, *Proceedings of the 14th Asian Conference on Computer Vision*, Springer, pp. 611–619. https://doi.org/10.1007/978-3-030-20893-6_19

[14] C. Godard, O. M. Aodha and G. J. Brostow (2017) Unsupervised Monocular Depth Estimation with Left-Right Consistency, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 6602–6611. https://doi.org/10.1109/CVPR.2017.699

[15] J. Bromley, I. Guyon, Y. LeCun, E. Sackinger, and R. Shah (1994) Signature verification using a siamese time delay neural network, *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 737–744. https://doi.org/10.1142/9789812797926_0003

[16] G. Koch, R. Zemel, and R. Salakhutdinov (2015) Siamese neural networks for one-shot image recognition, *Proceedings of International Conference on Machine Learning*.

[17] J. P. Yusiong and P. Naval (2019) AsiANet: Autoencoders in Autoencoder for Unsupervised Monocular Depth Estimation, *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, IEEE, pp. 443–451. https://doi.org/10.1109/WACV.2019.00053

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli (2004) Image quality assessment: from error measurement to structural similarity, *IEEE Transactions on Image Processing*, IEEE, vol. 13, no. 4, pp. 600–612. https://doi.org/10.1109/tip.2003.819861

[19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu (2015) Spatial transformer networks, *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 2017–2025.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele (2016) The cityscapes dataset for semantic urban scene understanding, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[21] Geiger, P. Lenz, and R. Urtasun (2012) Are we ready for autonomous driving? The kitti vision benchmark suite, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

[22] D. Eigen, C. Puhrsch and R. Fergus (2014) Depth map prediction from a single image using a multi-scale deep network, *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 2366–2374.

[23] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.* (2016) Tensorflow: a system for large-scale machine learning, *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, USENIX Association, pp. 265–283.

[24] D. Kingma and J. Ba (2015) Adam: A method for stochastic optimization, *Proceedings of the International Conference on Learning Representations*.

[25] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe (2017) Unsupervised learning of depth and ego-motion from video, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 6612–6619. https://doi.org/10.1109/CVPR.2017.700

[26] R. Mahjourian, M. Wicke, and A. Angelova (2018) Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 5667–5675. https://doi.org/10.1109/CVPR.2018.00594

[27] Z. Yin and J. Shi (2018) GeoNet: Unsupervised learning of dense depth, optical flow and camera pose, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1983–1992. https://doi.org/10.1109/CVPR.2018.00212

# Predicting the Causal Effect Relationship Between COPD and Cardio Vascular Diseases

Debjani Panda, Satya Ranjan Dash and Ratula Ray
Kalinga Institute of Industrial Technology, Bhubaneswar, India
E-mail: PANDAD@indianoil.in, sdashfca@kiit.ac.in, and ratularay@gmail.com

Shantipriya Parida (corresponding author)
Idiap Research Institute, Martigny, Switzerland
E-mail: shantipriya.parida@idiap.ch

**Keywords:** COPD, congestive heart failure, feature importance, ROC

**Received:** March 14, 2020

*Coronary Obstructive Pulmonary Disease (COPD) is one of the critical factors that are affecting the health of the population worldwide and in most cases affects the patient with cardiovascular diseases and their mortality. The onset of COPD in a patient in most of the cases affects him/her with cardiovascular disease and the management of the disease becomes more complex for medical practitioners to handle. The factors affecting COPD and cardiovascular disease in patients are most of the time, concurrent, and are responsible for their mortality. The list of factors and their underlying causes have been identified by experts and are treated with utmost importance before the patient suffers from an emergency condition and its management becomes even more difficult.*

*This paper discusses the need to study COPD and the factors affecting it to avoid cardiovascular deaths. The dataset used for the study is a novel one and has been collected from a Government Medical College, for study and experimentation. Classification methods like Decision Trees, Random Forest (RF), Logistic Regression (LR), SVM (Support Vector Machine), KNN (K-Nearest Neighbours), and Naïve Bayes have been used and Random Forests have given the best results with 87.5% accuracy. The importance of the paper is in the attempt to infer important links between the associated features to predict COPD. To the best of our knowledge, such an attempt to infer the interrelation between cardiac disease and COPD using Machine Learning classifiers has not been made yet. The paper focuses on determining the important correlation between the associated features of COPD and compare different supervised classifiers to check their prediction performance. Coronary Pulmonate, Age, and Smoking have shown a strong correlation with the presence of COPD and the performance analyses of the classifiers have been shown using the ROC (Receiver Operating Characteristic) curve.*

*Povzetek: Več metod strojnega učenja je bilo uporabljenih na povezovanju učinkov pljučnih in srčno-žilnih bolezni.*

## 1 Introduction

The concern for deaths due to COPD and Cardiovascular diseases is increasing worldwide in an exponential manner. Experts have identified a causal effect relationship between these two diseases where the presence of one determines the onset of the other or vice versa. COPD occurs in people complaining about severe difficulty in breathing or arrhythmia (irregular heartbeats). The presence of COPD in patients mostly makes them vulnerable to cardiovascular diseases and their mortality. It has also been observed that a patient suffering from cardiovascular disease also complains about COPD. This paper reveals the factors to be considered for patients with COPD for determining whether he is suffering from cardiovascular disease or not. COPD is characterized by obstruction in the air passages, which persist over a significant amount of time. It typically refers to bronchial asthma, bronchitis, and emphysema. Bronchial asthma is an allergic reaction that affects the respiratory tract, caused due to significant amounts of histamine in the blood. Bronchitis is defined as the inflammation of the bronchi caused due to infection. Emphysema is the inflation of bronchioles or alveolar sacs in the lungs. The common factors affecting the patients with COPD and Heart Disease have been shown in our below mentioned in Figure 1.

### 1.1 Symptoms of COPD

The common symptoms include coughing, wheezing, difficulty in breathing mainly during exhalation, excess mucous discharge, fatigue, pressure in the chest, anxiety, and loss

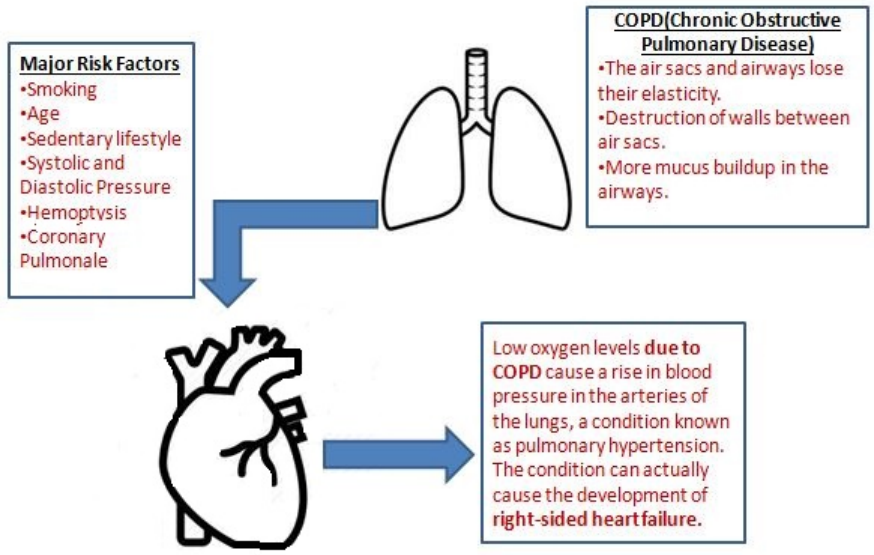



Figure 1: Causal Effect Relationship of heart disease and COPD.

in muscle tissue and weight.

### 1.2 Complications of COPD

People with COPD are more prone to fatal diseases such as pneumonia, pneumothorax (lung collapse), osteoporosis, edema, enlargement of liver and cor pulmonale (right side of the heart), diabetes, sleep apnea (repeated starting and stopping of breathing during sleep), stroke, high blood pressure, arrhythmia (irregular heartbeat), and heart failure. There has been evidence that patients with COPD [24] have a higher risk of Myocardial Infarction (MI) and it becomes worse when it is not properly managed in hospitals and adequate treatment is not given. Studies have revealed that factors like smoking contribute to around 3.8 % to 16 % of patients suffering from COPD and Congestive Cardiac Failure [7, 25]. Cardiovascular and COPD have comorbidities [17] like diabetes, smoking, hypertension, atria fibrillation, Congestive heart failure, and several other factors.

There is a necessity to study the factors that correlate COPD and Cardiovascular diseases [13, 9] so that their underlying factors can be identified and help in treating the patient on time to avoid premature deaths. COPD in most patients causes cardiovascular deaths, and it remains a challenge to identify its occurrence and treat the patient on time.

In this paper, we have used the real patient data from the Srirama Chandra Bhanja Medical College and Hospital [1], Cuttack, Odisha, India. The data was collected for 200 Patients by the Regional Medical Research Center (RMRC), Bhubaneswar [2]. RMRC is a medical Research Institute that collects patient's data for research and study of ailments. RMRC is the authorized body of Government to carry out research activities on available medical data of patients from different organizations. The data set used in our study was contributed by RMRC for research purposes. The data was collected by RMRC through questionnaires and from test reports of the patients with the consent of the patients. The raw data was pre-processed to drop identification labels of patients and missing values were imputed. The data was then split to train the classifiers. The experiment uses the heat map to identify the most important factors which affected the patients with COPD. Our paper reveals important factors like age, Coronary Pulmonate, and smoking as major contributing factors that are highly responsive to cause COPD in patients. The other factors include the systolic and diastolic pressure of the patients. Figure 2 shows the pipeline of the work that has been carried out in this paper.

## 2 Literature survey

Various supervised ML classifiers have been used for the prediction of health conditions for a long time. Works have been done previously to study the correlation between COPD and related risk factors. We have attempted to consolidate the most relevant works in this field that have been carried out in the past. Rabe et al. [21] establishes the strong connection between COPD and Cardio-Vascular diseases. Whenever a patient is suffering from either or both the diseases there are pathophysiological changes in the body which includes inflamed lungs and heart. The focus has been on suggesting various treatments to be administered to the patients suffering from COPD who have

[1] http://scbmch.nic.in/
[2] http://www.rmrcbbsr.gov.in/

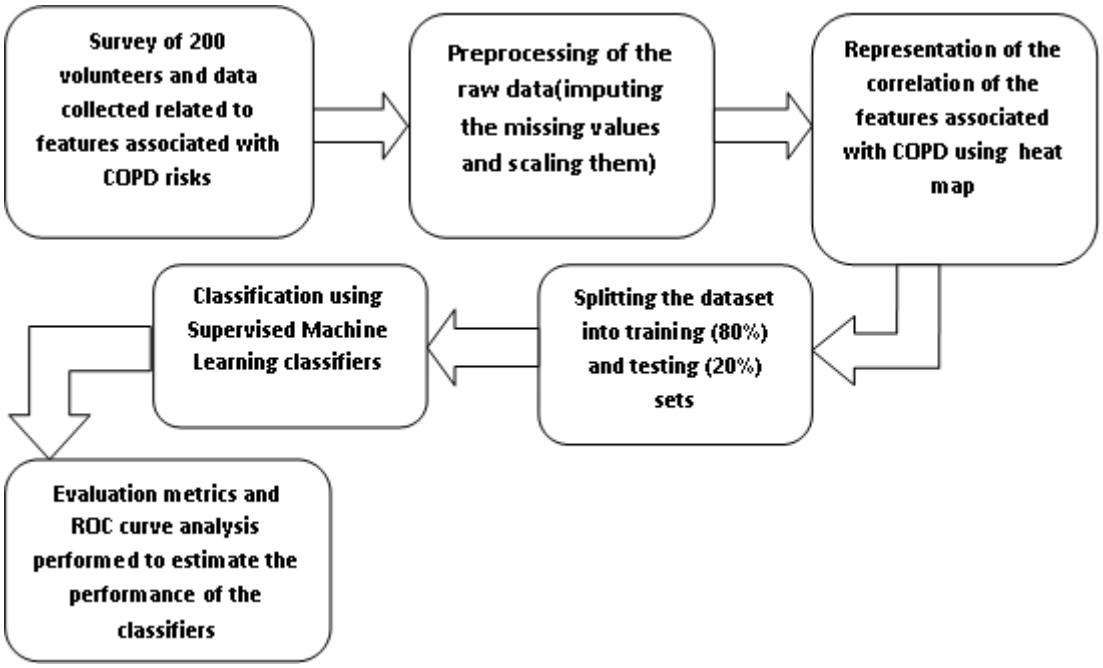



Figure 2: The figure highlights the workflow of the paper where the features associated with COPD are classified after preprocessing of the crude dataset to detect the presence of the disease. Also, the supervised classification and performance analysis of the different classifiers, using ROC analysis and other evaluation metrics are important prediction measures of the processed dataset.

these underlying heart diseases. The critical factors that have been identified to coexist with COPD and CVD are smoking, age, diabetes, and a sedentary lifestyle. The paper suggests the use of Beta-blockers for treating patients with Cardio Vascular ailments that do not have much interference with COPD drugs. Cazzola et al. [2] has suggested effective ways to manage the two interrelated deadly diseases COPD and CVD which coexist in the patients. The prime objective is to reduce the COPD causing symptoms by first treating the swollen lungs and breathlessness of the patients. Suggested drugs to reduce the urge to smoking, bronchodilators, and inhaled corticosteroids are used in the majority of the cases to treat COPD. The use of Angiotensin-converting enzyme inhibitors, angiotensin II type 1 receptor blockers, statins, antiplatelet drugs, or $\beta$-adrenoceptor blockers have been proved to be beneficial in treating patients with CVD which has effectively reduced COPD deaths in patients and reduced hospitalization for them. Holm et al. [10] focuses on identifying a genetic deficiency Alpha-I Antitrypsin deficiency and its impact on COPD with growing age. With growing age, the effect of psychological and clinical conditions of patients has been studied with COPD arising due to genetic deficiency (AATD). 468 individuals were considered for study with the genetic deficiency within varying age groups of 32 – 84 years. The individuals who were having severe AATD were found to be at greater risk of suffering from COPD. AATD is the genetic cause for the onset of COPD and this deficiency also aggravates smoking. The patients were studied for two years and from the study, it was observed that the younger generation was prone to anxiety, depression, and health issues regardless of their relationship status.

Fukuchi [8] significantly focuses on the growing age of individuals which is the factor of consideration for patients affected with COPD. Three models have been studied in which the animals were prematurely aged. Their lungs did not have any pathological changes like naturally aged lungs and were consistent in their function even after premature aging. The author has tried to state that the abnormal functioning of the lungs due to the increase in the size of the air spaces is not related to increasing age and the relationship between age and COPD is misleading. It has been suggested to further investigate the accelerated aging of lungs may be a factor to cause COPD but directly is not the cause.

The increased rate of smoking has also been greatly affecting patients and has been identified as the major cause of COPD. Laniado-Laborín [12] establishes the fact that smoking is the most important causal factor for patients suffering from COPD. The reduction of smoking in patients has been identified as a successful treatment for COPD patients. Different types of therapies i.e. both pharmacological and behavioral therapies have been suggested for stopping the progression of COPD in patients by controlling their smoking habits. However, studies have also suggested that the patients who refrained themselves from smoking after one year of follow-up were very low. Studies have also revealed that pharmacological therapies are more effective than placebo and 25-30% of people have abstained from smoking after taking these therapies. But, still smoking cessation remains a major challenge in the world and

patients continue to suffer from COPD due to their smoking habits. Khan et al. [11] describes in their paper the after-effects of smoking and how this creates abnormalities in lung function. It is responsible for the thickening of airways, dilation of air spaces with abnormal distension of alveoli. It is observed that almost all cigarette smokers have inflammation in their lungs. It has been summarized that tobacco inhalation active or passive results in abnormal inflammation, leads to tissue-damaging oxidants, a reduced level of antioxidants (for self-protection), and induced cell death.

Work has also been done to bring out the association of cor pulmonale with dysfunction of lungs by Shujaat et al. [23]. The earlier fact that the right ventricular dysfunction of the heart is due to enlargement of the tissue results because of left ventricular dysfunction adds up to the information that the right ventricle of the heart has an underlying cause of the malfunction of the lungs and its size. They have identified Pulmonary Hypertension as the underlying cause of right ventricle dysfunction resulting in heart failure. Cor pulmonale with pulmonary hypertension occurs due to various reasons and to treat the patients with the disease several treatments like inhalation of Nitric Oxide, usage of diuretics to remove excess water from the lungs and heart, giving a pulmonary vasodilator like sildenafil, reduction in hematocrit, and surgery for reducing lungs size have been suggested to reduce the cardiac arrests in patients suffering from COPD.

Quint [20] focuses on the fact that patients suffering from COPD have a higher risk of suffering from cardiovascular diseases. The author summarizes delayed identification of disease, late treatment given for reperfusion of STEMI (ST-Elevation Myocardial Infarction), and use of angiography after in-STEMI as causes of a gap for mortality of COPD patients suffering from Myocardial Infarction. Troponin has been identified as the direct indicator in patients suffering from COPD. The results revealed that the higher the presence of Troponin in cardiac patients, the longer they stay in hospitals and have less chance of survival.

The link to various diseases associated with COPD has also been previously studied. Feary et al. [6] has tried to summarize the diseases that are associated with inflammation in the lungs which include cardiovascular diseases and diabetes mellitus as the most common associated diseases. The study determines quantitatively the effect of cardiovascular diseases for patients suffering from COPD. The data of patients have been analyzed with logistic regression with multiple variables and Cox regression. The presence of cardiovascular diseases is found to be more in the young age group of COPD affected patients after considering several factors such as smoking and age strata of patients. It has been found that in most cases, patients suffering from COPD are already affected by myocardial infarction and diabetes.

Roever et al. [22] discussed the factors responsible for COPD and cardiovascular diseases. The authors have identified various factors such as sedentary lifestyle, systemic inflammation, improper function of skeletal muscle, etc for being responsible for cardiovascular diseases and which again becomes the major reason for patients suffering from COPD. The diseases like diabetes, hypertension, a metabolic syndrome that arises due to smoking, arterial fibrillation, Vitamin D deficiency, Congestive cardiac failure are among the several factors affecting patients suffering from heart disease. These patients are found to be vulnerable to COPD and in most cases are affected which is the ultimate factor for causing deaths. COPD patients mostly die of strokes and cardiovascular-related diseases have been identified as the major cause affecting these patients.

Esteban et al. [5] mentioned the factors responsible for the worsening of COPD. Using machine learning an early prediction system is designed which will warn about chronic obstructive pulmonary disease in three levels: red, yellow, and green. The model used Random forests for predicting the condition of COPD in patients. The attributes of the data set were obtained from the daily activities and questionnaires from the patients. The model was trained and tested using 10-fold cross-validation for improving the performance. The model achieved a ROC curve of 0.87 for forecasting whether a patient will suffer from COPD worsening within the next three days.

Peng et al. [19] developed a model for predicting acute illness in patients affected with Chronic Obstructive Pulmonary disease. The 28 important features were selected from watches, sphygmomanometers, thermometers, and routine clinical tests. They identified 410 records from the hospital database for the study and the trained: test ratio varied from 90:10 to 50:50. The best results were obtained with a split in the 80:20 ratio. The model used C4.5 and C5.0 decision trees for classification of the disease. ID3, CART, and C 5.0 classifiers have been compared to find out the best model. C5.0 with 80:20 train and test split gave 80.3% accuracy.

Xie et al. [26] tried to relate physiological homeostasis and the onset of COPD exacerbation. A regression model is built to study the patterns of variables extracted from patients and are evaluated longitudinally for all records to classify the nature of risk the patient is subjected to. The data set was obtained from a hospital in Sydney using TeleMedCare Health Monitor which included the attributes: weight, diastolic and systolic blood pressure (DBP, SBP, heart rate (HR), SpO2, and temperature. The model used Logistic regression with double loop 10- fold cross-validation and has yielded 79.27% accuracy with AUC of 0.84.

| Attribute Name | Allowed Values | Description |
|---|---|---|
| Gender | 1=Female, 0=Male | Gender of the patient |
| Lifestyle | 1=Sedentary, 0=Active | Type of lifestyle which is adopted by the patient |
| Literacy | 1:Illiterate, 0=Literate | The patients is literate or illiterate |
| Family History | 1=Has family history, 0=does not have family history | Whether the patient has any family history of COPD. |
| Weight | | The actual weight of the patient in Kgs. |
| Systolic Pressure | | It's the blood pressure measures the pressure in your blood vessels when your heart beats |
| Diastolic Pressure | | Blood pressure while the heart rests. |
| Alcohol | 1=drinks, 0=does not drink | The patient consumes alcohol or not. |
| Smoking: | 1=smokes, 0= does not smoke | The patient smokes or is a non-smoker. |
| Age | | Current age if the patient. |
| Col Pulmonate | 1=Present, 0=absent | Right side heart failure due to pulmonary hypertension. |
| Hemoptysis: | 1=Present, 0=absent | Patient is coughing out with blood or not. |
| Type | 1=COPD detected, 0= COPD not detected | Patient is affected with COPD or not. |

Table 1: Data Set Detail Description

# 3 Materials/methods

## 3.1 Methods

### 3.1.1 Naïve Bayes classification

Naive Bayes classifiers are a throng of classification algorithms based on Bayes' Theorem. This classification technique makes an assumption of independence among predictors. In layman's terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayes' Thoerem [14] is based on probability theory:

$$P(A|B) = P(A)P(B|A)/P(B), \tag{1}$$

where $P(A|B)$ is how often A happens given that B happens,
$P(B|A)$ is how often B happens given that A happens,
$P(A)$ is how likely A is on its own,
$P(B)$: is how likely B is on its own.

### 3.1.2 Support vector machine (SVM)

SVM is a set of learning methods that are supervised and used for classification and regression. An SVM model is a representation of the data as examples in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. This is done by identifying a hyperplane [27] which separates the classified data with maximum space between them and is determined in such a fashion that most of the points of one category fall on one side of the plane. SVM determines the best-fitted plane.

### 3.1.3 Decision trees

Decision Tree is a tool that represents nodes of the tree and helps take decisions depending upon the inputs of a node. It helps in giving a pictorial presentation of the consequences of a certain condition. It is used in classification and regression. Here the nodes represent the data [18] and not the decisions. Here a threshold value has to be given after which the algorithm will terminate. It is left with some points which could not be classified and this is called Gini impurity.

### 3.1.4 KNN classification

K Nearest Neighbor algorithm is a non-parametric method used for classification and regression [1]. This classification technique calculates the distance of a point with coordinates (x, y) from its neighbors. For example, when there are two sets of points in the space and a new point has to be plotted in the area, then the question is where it should be plotted and with which region to determine its basic characteristics that satisfy the classification correctly. The Euclidean distance is calculated from the point from its neighbors and finally, it is positioned in the area which is closest to its neighboring points.

### 3.1.5 Logistic regression

It is a statistical tool that is used for making decisions on the binary output of the testing condition. In the Linear model the equation used is:

$$y = b_0 + b_1(x), \tag{2}$$

whereas logistic regression uses the equation:

$$P = \frac{1}{1 + e^{-(b_o + b_1 x)}} \tag{3}$$

In the logistic regression, the constant $(b_0)$ moves the curve left and right and the slope $(b_1)$ defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio:

$$\frac{p}{1-p} = exp^{(b_0 + b_1 x)} \tag{4}$$

### 3.1.6 Random forest

The random forest classifier [15] creates a set of decision trees instead of a single one which is generated by randomly taking one seed from a selected subset of the data of the training set. Subsequently, it determines the average of the results from different decision trees to determine the class to which the object belongs to. It is an ensemble method [4, 3] for classification as well as a regression that operates by constructing an assembly of decision trees at training time and finding out the class that is the most suitable for its outcome depending upon its predicted values.

# 4 Experiment

The following methodology has been adapted for the collection, study, and experimentation of the data which has been contributed by RMRC.

1. Carrying out of the survey of the 200 patients who volunteered to share their health cards and reports to carry out the research work.
2. Taking medical experts' opinions to identify which factors are related to the COPD occurrence.
3. Pre-processing of the raw data(imputing the missing values and dropping the features which are not directly relevant to the outcome).
4. Finding the correlation between all the risk factors with the presence or absence of COPD in the patient samples using heat-map. In this process, we can find the most relevant features associated with the outcome.
5. Splitting the entire dataset into training (80%) and testing sets (20%).
6. Using Supervised learning classifiers to classify the dataset in giving the prediction of the outcome (Gaussian Naïve Bayes, SVM, Decision Trees, Logistic Regression, Random Forest and KNN )

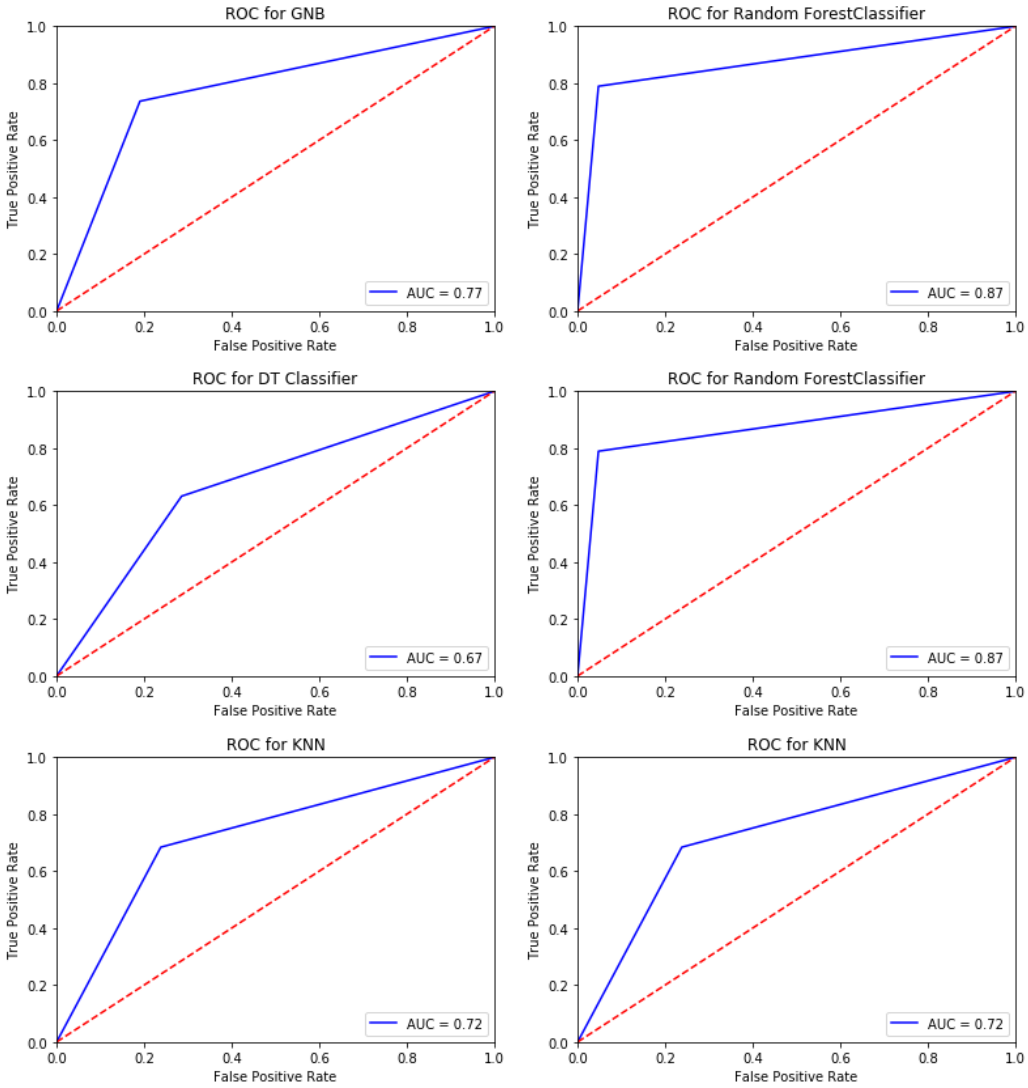



Figure 3: ROC Analysis of Gaussian Naive Bayes,Random Forest, Decision Trees, Logistic Regression, KNN and SVM Classifiers in which Random forest has given the best curve .The blue line aligning towards left represents the true positive rate of the predictive classifiers and the red line demarcates the threshold. The quality of the performance of the classifiers depends upon the nature of the blue curve.

# 5 Results

Our work has been implemented in Python to study the behavior of the supervised algorithms. The original data set has been collected from SCB Medical College through a survey conducted by RMRC. The dataset is novel and has been used for study only. The raw data was collected for 200 patients of Government Medical College to identify the critical connection between COPD and heart disease. Results of the six Classification Algorithms have been noted in Table 2 with their confusion matrix/accuracy in predicting the test samples.

| Classifier | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| GNB | 77.5 | 80.95 | 77.27 | 0.77 |
| RF | 87.5 | 95.23 | 90.90 | 0.87 |
| DT | 67.5 | 71.43 | 68.18 | 0.67 |
| LR | 82.5 | 85.71 | 81.81 | 0.82 |
| KNN | 72.5 | 76.19 | 72.72 | 0.72 |
| SVM | 77.5 | 80.95 | 77.27 | 0.77 |

Table 2: Performance Comparison of the Classifiers

The identification of important features [16] has been done using the heat map attached in Figure 4 and the factors which are mostly responsible for causing COPD have been determined as Cor Pulmonale, Age, and Smoking. These factors are then studied with several algorithms to analyze their performance in terms of accuracy, precision, and recall.

The performance of the classifiers have been studied by plotting the ROC for all methods as shown in Figure 3. The results show the best curve has been given by a random forest classifier which has the area under the curve 0.87. The true positives obtained for Random Forest are more than any other classifiers used. The Logistic Regression and SVM Classifiers have also given good results with AUC 0.77, which can be explicitly seen from the ROC curves.

From the data collected, the best results were produced by Random forest classifier(accuracy: 87.5%) with a Precision of 95.23% and a recall score is 90.90%.

# 6 Discussion

Various works have been carried out to identify the symptoms and factors affecting the health of COPD patients. A strong connection is found to be existing in patients with COPD, who are also affected by Cardiovascular disease [21, 2, 23]. Our paper also has stressed the explicit fact that these two diseases are interrelated and the onset of one disease causes the other disease to impact the patient soon.

The factors affecting the condition of COPD have been identified to be smoking, Cor pulmonate, age, and diabetes [23, 20, 6, 22], and other lifestyle-related factors. Similar facts have been established in our work, where the important factors obtained from the heatmap denote smoking, coronary pulmonate, and age. Even to some extent literacy status and lifestyle played a vague role.

The papers which were used to study the performance of models for predicting COPD gave an accuracy of 79-80% [5, 19, 26] with max. ROC of 0.87. Our model has given superior results than other models and has given the highest accuracy of 87.5% with a ROC of 0.87. There is an improvement of almost 7% in the prediction of COPD patients using our suggested model.

# 7 Conclusion

Random forests being an ensemble classifier have given the best results for predicting the condition of a patient affected with COPD. The most important factors which have been identified as the causal ones include age, smoking, and Cor pulmonale. In addition to these Systolic and diastolic pressure also have been identified to have an impact on the underlying disease. Cor pulmonale is an abnormal enlargement of the heart due to infection in the lungs or any blood vessels. The above facts conclude that COPD is related to heart disease and gets worsened with the systolic and diastolic pressure of the patient. The systolic and diastolic pressure gets affected when the person is suffering from heart disease. With our data set, we have identified that these two diseases coexist, or both are interrelated to each other. Our work has been done with a novel data set of 200 patients. The work illustrates the fact that classification methods can be used to find the relationship between various diseases and their occurrences. The study can be extensively made for more number of patients so that the model behavior can be the same for new experimental case studies. The classification methods can be combined with optimization techniques for better prediction accuracy. Our model can be used by medical experts to determine heart disease well in advance and the symptoms of COPD can be interpreted by the model to predict occurrences of fatal diseases.

# References

[1] G. E. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004. https://doi.org/10.1145/1007730.1007735.

[2] M. Cazzola, L. Calzetta, B. Rinaldi, C. Page, G. Rosano, P. Rogliani, and M. G. Matera. Management of chronic obstructive pulmonary disease in patients with cardiovascular diseases. *Drugs*, 77(7): 721–732, 2017. https://doi.org/10.1007/s40265-017-0731-3.

[3] S. R. Dash and S. Dehuri. Comparative study of different classification techniques for post operative patient dataset. *International Journal of Innovative Re-*

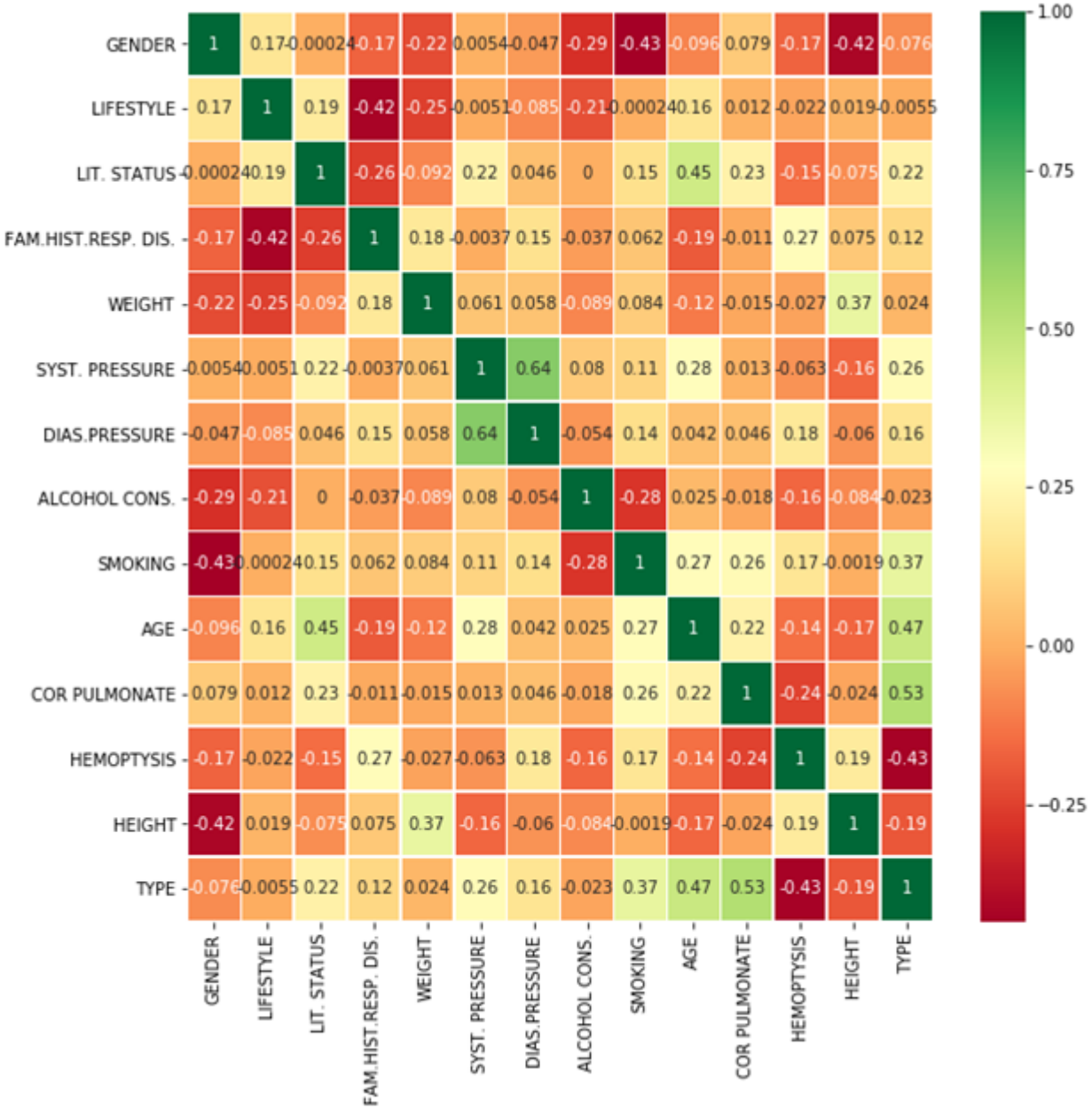



Figure 4: Heat Map determines Cor Pulmonale, age, and Smoking as Important Features. The Heat Map provides an important correlation between different features that are important about the classification process and assigns a degree of correlation, which is projected through the color mapping

*search in Computer and Communication Engineering*, 1(5):1101–1108, 2013.

[4] S. R. Dash and R. Ray. Predicting seminal quality and its dependence on life style factors through ensemble learning. *International Journal of E-Health and Medical Communications (IJEHMC)*, 11(2):78–95, 2020. https://doi.org/10.4018/ijehmc.2020040105.

[5] C. Esteban, J. Moraza, F. Sancho, M. Aburto, A. Aramburu, B. Goiria, A. Garcia-Loizaga, and A. Capelastegui. Machine learning for copd exacerbation prediction, 2015. https://doi.org/10.1183/13993003.congress-2015.oa3282.

[6] J. R. Feary, L. C. Rodrigues, C. J. Smith, R. B. Hubbard, and J. E. Gibson. Prevalence of major comorbidities in subjects with copd and incidence of myocardial infarction and stroke: a comprehensive analysis using data from primary care. *Thorax*, 65(11): 956–962, 2010. https://doi.org/10.1136/thx.2009.128082.

[7] F. M. Franssen and C. L. Rochester. Comorbidities in patients with copd and pulmonary rehabilitation: do they matter?, 2014. https://doi.org/10.1183/09059180.00007613.

[8] Y. Fukuchi. The aging lung and chronic obstructive pulmonary disease: similarity and difference. *Proceedings of the American Thoracic Society*, 6(7): 570–572, 2009. https://doi.org/10.1513/pats.200909-099rm.

[9] K. Ghoorah, A. De Soyza, and V. Kunadian. Increased cardiovascular risk in patients with chronic obstructive pulmonary disease and the potential mechanisms linking the two conditions: a review. *Cardiology in review*, 21(4):196–202, 2013. https://doi.org/10.1097/crd.0b013e318279e907.

[10] K. E. Holm, M. R. Plaufcan, D. W. Ford, R. A. Sandhaus, M. Strand, C. Strange, and F. S. Wamboldt. The impact of age on outcomes in chronic obstructive pulmonary disease differs by relationship status. *Journal of behavioral medicine*, 37(4): 654–663, 2014. https://doi.org/10.1007/s10865-013-9516-7.

[11] S. Khan, P. Fell, and P. James. Smoking-related chronic obstructive pulmonary disease (copd). *Diversity and Equality in Health and Care*, 11(3-4):267–271, 2014.

[12] R. Laniado-Laborín. Smoking and chronic obstructive pulmonary disease (copd). parallel epidemics of the 21st century. *International journal of environmental research and public health*, 6(1): 209–224, 2009. https://doi.org/10.3390/ijerph6010209.

[13] J. D. Maclay and W. MacNee. Cardiovascular disease in copd: mechanisms. *Chest*, 143(3):798–807, 2013. https://doi.org/10.1378/chest.12-0938.

[14] R. Mitchell, J. Michalski, and T. Carbonell. *An artificial intelligence approach*. Springer, 2013.

[15] D. Panda and S. R. Dash. Predictive system: Comparison of classification techniques for effective prediction of heart disease. In *Smart Intelligent Computing and Applications*, pages 203–213. Springer, 2020. https://doi.org/10.1007/978-981-13-9282-5_19.

[16] D. Panda, R. Ray, A. A. Abdullah, and S. R. Dash. Predictive systems: Role of feature selection in prediction of heart disease. In *Journal of Physics: Conference Series*, volume 1372, page 012074. IOP Publishing, 2019. https://doi.org/10.1088/1742-6596/1372/1/012074.

[17] A. I. Papaioannou, K. Bartziokas, S. Loukides, S. Tsikrika, F. Karakontaki, A. Haniotou, S. Papiris, D. Stolz, and K. Kostikas. Cardiovascular comorbidities in hospitalised copd patients: a determinant of future risk? *European Respiratory Journal*, 46(3): 846–849, 2015. https://doi.org/10.1183/09031936.00237014.

[18] B. N. Patel, S. G. Prajapati, and K. I. Lakhtaria. Efficient classification of data using decision tree. *Bonfring International Journal of Data Mining*, 2 (1):06–12, 2012. https://doi.org/10.9756/bijdm.1098.

[19] J. Peng, C. Chen, M. Zhou, X. Xie, Y. Zhou, and C.-H. Luo. A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. *Scientific reports*, 10 (1):1–9, 2020. https://doi.org/10.1038/s41598-020-60042-1.

[20] J. Quint. The relationship between copd and cardiovascular disease. *Tanaffos*, 16(Suppl 1):S16–S17, 2017.

[21] K. F. Rabe, J. R. Hurst, and S. Suissa. Cardiovascular disease and copd: dangerous liaisons? *European Respiratory Review*, 27(149), 2018. https://doi.org/10.1183/16000617.5057-2018.

[22] L. Roever et al. Translational medicine. 2015.

[23] A. Shujaat, R. Minkin, and E. Eden. Pulmonary hypertension and chronic cor pulmonale in copd. *International journal of chronic obstructive pulmonary disease*, 2(3):273–282, 2007.

[24] D. D. Sin and S. P. Man. Chronic obstructive pulmonary disease as a risk factor for cardiovascular morbidity and mortality. *Proceedings of the American Thoracic Society*, 2(1):8–11, 2005. https://doi.org/10.1513/pats.200404-032ms.

[25] A. Undas, P. Kaczmarek, K. Sladek, E. Stepien, W. Skucha, M. Rzeszutko, I. Gorkiewicz-Kot, and W. Tracz. Fibrin clot properties are altered in patients with chronic obstructive pulmonary disease. *Thrombosis and haemostasis*, 102(12):1176–1182, 2009. https://doi.org/10.1160/th09-02-0118.

[26] Y. Xie, S. J. Redmond, M. S. Mohktar, T. Shany, J. Basilakis, M. Hession, and N. H. Lovell. Prediction of chronic obstructive pulmonary disease exacerbation using physiological time series patterns. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6784–6787. IEEE, 2013. https://doi.org/10.1109/embc.2013.6611114.

[27] Y. Yang, J. Li, and Y. Yang. The research of the fast svm classifier method. In *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 121–124. IEEE, 2015. https://doi.org/10.1109/iccwamtip.2015.7493959.

# Probabilistic Weighted Induced Multi-Class Support Vector Machines for Face Recognition

Aniruddha Dey
Department of Information Technology, MAKAUT, Salt Lake, Kolkata, India
E-mail: anidey007@gmail.com

Shiladitya Chowdhury
Department of Master of Computer Application, Techno India, Kolkata, India
E-mail: dityashila@yahoo.com

**Keywords:** face recognition, weighted multi-class SVM, optimal separating hyperplane, probabilistic method.

**Received:** April 29, 2020

*Abstract: This paper deals with a probabilistic weighted multi-class support vector machines (WMSVM) for face recognition. The support vector machines (SVM) has been applied to many application fields such as pattern recognition in last decade. The support vector machines determine the hyperplane which separates largest fraction of samples of the similar class on the same side. The SVM also maximizes the distance from the either class to the separating hyperplane. It has been observed that in many realistic applications, the achieved training data is frequently tainted by outliers and noises. Support vector machines are very sensitive to outliers and noises. It may happen that a number of points in the training dataset are misplaced from their true position or even on the wrong side of the feature space. The weighted support vector machines are designed to overcome the outlier sensitivity problem of the support vector machines. The main issue in the training of the weighted support vector machines algorithm is to build up a consistent weighting model which can imitate true noise distribution in the training dataset, i.e., reliable data points should have higher weights, and the outliers should have lower weights. Therefore, the weighted support vector machines are trained depending on the weights of the data points in the training set. In the proposed method the weights are generated by probabilistic method. The weighted multi-class support vector machines have been constructed using a combination of the weighted binary support vector machines and one-against-all decision strategies. Numerous experiments have been performed on the AR, CMU PIE and FERET face databases using different experimental strategies. The experimental results show that the performance of the proposed method is superior to the multi-class support vector machines in terms of recognition rate.*

*Povzetek: Opisana je metoda podpornih vektorjev za prepoznavanje obrazov.*

## 1 Introduction

The SVM can be considered as an estimated implementation of the structural risk minimization method [1]. In 1998, Vapnik first devised the SVM to address the pattern classification and recognition problem [2]. The objective of the support vector machines is to determine the hyperplane that divides largest fraction of images in the related class on the same adjacent, whereas maximizing the space from the both class to the separating hyperplane. This separating hyperplane is known as optimal separating hyperplane (OSH). The OSH minimizes the misclassification risk. It may be noted that in many realistic applications, some training data points are placed far away from the accurate position or even on the wrong side of the feature space. These data points are called outliers. In general, the training dataset is severely affected by the outliers and different kind of noises. The SVMs are actual sensitive to outliers and different kind of noises. Therefore, in the training phase, the outliers with large Lagrangian coefficient can become a support vector [3]. In the past few decades, wide ranges of techniques have been introduced by several researchers to solve the aforementioned bottleneck of the SVM. Zhang [4] proposed central SVM (CSVM) in which class centres are used to build the support vector machines. For each training data point, the adaptive margin SVM (AMSVM) training algorithm [5] depends on the utilization of the adaptive margins. Song et al. [6], [7] proposed a robust SVM (RSVM) in which to generate an adaptive margin, the space between centre of every class of the training sample and the data point is computed. But this method has a drawback because it is very difficult to tune the penalty parameter. The method uses the averaging method which is partly sensitive for outliers and noises. Authors in [8] and [9] proposed fuzzy SVM (FSVM) to eliminate the outlier sensitivity problem. To moderate the effect of outliers, the method applies the fuzzy membership's values to the training data. Membership function selection is main drawback for the FSVM. Cao et al. [10] proposed the support vector novelty detector

(SVND) which detects the outliers more appropriately from the normal data points, and solve one-class classification problem.

Some new improvements on the support vector machines can be establish in the literature review. Quan *et al*. [11] established the weighted least squares support vector machine (WLS-SVM) local region algorithm. This algorithm calculates the nonlinear time series, as well as performs robust estimation for regression using the limited observations. In this method, there is a simple and effectual technique to model parameter selection based on the leave one-out cross-validation strategy. A weighting method on Lagrangian SVM (LSVM) is proposed by Hwang *et al*. [12]. This method deals with the imbalanced data classification problem. In this method, a weight parameter is added to the LSVM design. Therefore, the method can get better performance for the minority class with minimum control on classification performance of the majority class. Yu [13] proposed the asymmetric weighted least squares support vector machine (LSSVM) combined learning procedure. This methodology is based on the evolutionary programming (EP), and is used for software repository mining. A nonparallel plane classifier, namely, weighted twin support vector machines with local information (WLTSVM) is proposed by Ye *et al*. [14]. This method mines underlying similarity information within the samples as much as possible. Shao *et al*. [15] proposed the weighted Lagrangian twin support vector machines (WLTSVM) for the imbalanced data classification. Xanthopoulos *et al*. [16] suggested the weighted support vector machines for automated procedure checking and early error diagnosis. The robust LS-SVM (RLS-SVM) is proposed by Yang *et al*. [17], and the method is established on the truncated least squares loss function for classification and regression with noises. Zhang *et al*. [26] proposed an emotion recognition system based on facial expression images. In this work**,** the bi-orthogonal wavelet entropy is used to extract multi-scale features and the fuzzy multi-class support vector machine is used as classifier. More recently, Wang *et al.* offered a new intelligent emotion recognition system where stationary wavelet entropy are used to extract feature values and a single hidden layer feed forward neural network is employed as the classifier [27]. Aburomman and Reaz proposed ensemble classifiers are generated using the novel methods as well as the weighted majority algorithm (WMA) technique [28]. Some learning based discriminant analysis techniques have been suggested, such as local structure preserving discriminant analysis [29], Discriminant similarity and variance preserving projection [30] to abuse the label info contained in the data. Shi*et al.* established 3D face recognition method based on LBP and SVM.Hu and Cui proposed Digital image recognition based on Fractional order PCA-SVM coupling algorithm [32]. By improve the SVM gender classification accuracy using clustering and incremental learning suggested by Dagher and Azar [33].Kar*et al*. face expression recognition system based on ripplet transform type II and least square SVM [34].

In this study, the probabilistic weighted multi-class support vector machine is devised to address the outlier sensitivity problem. The main issue in the training samples of the weighted support vector machines algorithm is to improve a reliable weighting model which can reflect true noise distribution in the training data, i.e., reliable data points should have higher weights, and the outliers should have lower weights. Therefore, dissimilar weights are allocated to different data points. Therefore, as per relative importance of the data points in the training set, the training algorithm of the weighted SVM determines the decision surface. The probabilistic method is used to generate the weights of the proposed probabilistic weighted multi-class support vector machines training algorithm. These weights are incorporated with all data points of the training set. The weighted support vector machines training algorithm maximizes the margin of separation with the help of weights to prevent some points. In this work, the generalized two-dimensional Fisher's linear discriminant (G-2DFLD) technique is applied for feature extraction [18]. The extracted features are applied on the proposed probabilistic weighted multi-class support vector machines for training, classification and recognition. The empirical results on the AR, CMU PIE and FERET face database illustrate that the proposed probabilistic weighted multi-class support vector machines (WMSVM) perform better than the multi-class SVM, in terms of face recognition.

Rest of the paper is ordered as follows. The basic idea of the SVM is given in Section 2. The proposed weight generating scheme, based on the probabilistic method, is discussed in Section 3. Section 4 describes the weighted support vector machines. The weighted multi-class support vector machines are defined in Section 5. The simulation results on the AR, CMU PIE, and FERET face databases are described in Section 6. Section 7 contains the concluding remarks.

## 2 Revisited support vector machines

The support vector machines were developed for binary pattern classification problem [1 -3]. It has been seen that in case of pattern classification problem, the SVMs provide satisfactory performance. The basic idea of the binary-class SVMs [1- 3] is to split two classes by a hyperplane. This separating hyperplane is created from the available training samples. The support vector machines find the hyperplane that splits largest fraction of samples of the alike class on the similar side, while maximizing the space from the each class to the separating hyperplane. This separating hyperplane is known as optimal separating hyperplane (OSH). The OSH reduces the misclassification risk.

## 3 Weight generation by the probabilistic method

Although, the support vector machines are very powerful for solving classification problem, however, it has some limitations as it treats all the training data points of a

given class uniformly. It has been seen that, all the data points of the training set are not equally important for classification and recognition purpose in many real world application domains. This limitation of the support vector machines can be overcome by designing the weighted support vector machines. In the weighted support vector machines, each and every data points are treated separately according to their weights.

The main issue of the training algorithm of the weighted support vector machines is to develop a reliable weighting model which can reflect actual distribution in the training set. The reliable data points should have higher weights, and the outliers should have lower weights. Therefore, dissimilar weights are assigned to different data points. The decision surface generated by the weighted SVM training algorithm considers the relative significance of data points in the training set. The weights employed in the proposed probabilistic weighted multi-class support vector machines are generated by the probabilistic method.

Let the $c^{th}$ class has $N_c$ numbers of training samples. We consider the positive samples are belonging to class $y_1$ and negative samples are belonging to class $y_2$ to design the weighted SVM for $c^{th}$ class.

Let $P(y_j)$; $j \in 1,2$, defined the prior probability of the sample which is included in $y_j$ class. The prior probability of the sample belonging to $y_1$ class can be described as follows:

$$P(y_1) = \frac{N_c}{N} \tag{1}$$

Similarly, the prior probability of the sample belonging to $y_2$ class can be demonstrated as follows:

$$P(y_2) = \frac{N - N_c}{N} \tag{2}$$

Now for a positive training sample $\boldsymbol{x}_i$ the weight $a_i$ is illustrated as follows:

$$a_i = P(y_1|\boldsymbol{x}_i) = \frac{P(\boldsymbol{x}_i|y_1)\,P(y_1)}{P(\boldsymbol{x}_i|y_1)\,P(y_1)+P(\boldsymbol{x}_i|y_2)\,P(y_2)} \tag{3}$$

Similarly, in case of a negative training sample $\boldsymbol{x}_i$ the weight $a_i$ is generated as follows:

$$a_i = P(y_2|\boldsymbol{x}_i) = \frac{P(\boldsymbol{x}_i|y_2)\,P(y_2)}{P(\boldsymbol{x}_i|y_1)\,P(y_1)+P(\boldsymbol{x}_i|y_2)\,P(y_2)} \tag{4}$$

It is to be noted that $\varepsilon < a_i < 1$, and $\varepsilon(\varepsilon > 0)$ is sufficiently small. The term $P(y_j \mid \boldsymbol{x}_i)$; $j \in 1,2$ is called posterior probability, i.e., probability of the class is $y_j$ after we have performed measurement on the data $\boldsymbol{x}_i$. Similarly, the term $P(\boldsymbol{x}_i \mid y_j)$; $j \in 1,2$ is called conditional probability i.e., the probability that the class $y_j$ has the feature value $\boldsymbol{x}_i$. The equations (3) and (4) ensure that the lower weights are assigned to outliers or close to outliers.

Every measurement must be assigned to one of these two classes$y_1$ or $y_2$. Therefore,

$$\sum_{j=1}^{2} P(y_j \mid \boldsymbol{x}_i) = 1 \tag{5}$$

The posterior probability of the sample xi $(P(y_j \mid \boldsymbol{x}_i);\ j \in 1,2)$ is used as weight for designing the proposed probabilistic weighted multi-class support vector machines.

# 4 Weighted support vector machines

It has been seen that the training dataset is often tainted by outliers and noises in many real world applications. The support vector machines are very sensitive to outliers and noises. It may so happen that some patterns in the training set are outliers and misplaced far away from the true position or even on the wrong side of the feature space. During the training process, the outlier with large Lagrangian coefficient can become a support vector. The optimal hyperplane obtained by the support vector machines depends only on small part of the data points, i.e., support vectors. So, in presence of outliers, the decision boundary obtained by the support vector machines training algorithm deviate severely from the optimal separating hyperplane.

The weighted support vector machines are designed to address this issue. In weighted support vector machines, the data points of the training set are treated differently according to their weights. The training algorithm gives more effort to correctly classify more important data points (i.e., the data points with larger weights) while caring less effort to less important data points (i.e., the data points with lower weights, probably outliers).

Let $B$ be a set of labeled training samples associated with weights:

$$B = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^{N};\quad \boldsymbol{x}_i \in \Re^d;\quad y_i \in \{+1,-1\} \tag{6}$$

where, $\boldsymbol{x}_i$ is the input pattern for the $i^{th}$ training sample, $a_i$ is the weight assigned to $\boldsymbol{x}_i$, and $y_i$ is the class of the $\boldsymbol{x}_i$. In the proposed probabilistic weighted multi-class support vector machines, the weight is generated by the weight generating technique described in section 3.

To achieve better performance, the weighted support vector machines training algorithm maximizes the margin of separation. The optimal separating hyperplane in the case of weighted support vector machines minimizes the following function:

$$\Gamma(\boldsymbol{\omega}, \xi, a) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\sum_{i=1}^{N} a_i\xi_i \tag{7}$$

with constraints defined [1, 2].

In the optimization problem, the effect of the parameter $\xi_i$ is reduced by the small value of $a_i$. Therefore, the training algorithm of the weighted SVM considers the corresponding point ($\boldsymbol{x}_i$, $y_i$) as less significant for classification.

The solution to the optimization problem (7), subject to the constraints defined in [1, 2], is given by the saddle point of the following Lagrange function:

$$L(\boldsymbol{\omega}, b, \xi_i, \lambda) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\sum_{i=1}^{N} a_i\xi_i - \sum_{i=1}^{N}\lambda_i(y_i((\boldsymbol{\omega}^T \cdot \boldsymbol{x}_i)+b)-1+\xi_i) - \sum_{i=1}^{N}\gamma_i\xi_i \tag{8}$$

By expanding equation (8) term by term, the following equation is obtained.

$$L(\boldsymbol{\omega},b,\xi_i,\lambda)=\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}+C\sum_{i=1}^{N}a_i\xi_i-\sum_{i=1}^{N}\lambda_i y_i(\boldsymbol{\omega}^T\cdot\boldsymbol{x}_i)- b\sum_{i=1}^{N}\lambda_i y_i+\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_i\xi_i-\sum_{i=1}^{N}\gamma_i\xi_i \quad (9)$$

The Lagrange multipliers $\gamma_i$ are presented in equations (8) and (9) to ensure the non-negativity of slack variables $\xi_i$. At saddle point, the Lagrange function (8) has to be minimized with respect to $\boldsymbol{\omega}$, $b$, and $\xi_i$. It has to be also maximized with respect to $\lambda_i$ where, $0\le\lambda_i\le a_iC$.

We can convert the Lagrange function (8) into its corresponding dual problem as follows:

$$\max_{\lambda}\boldsymbol{\omega}(\lambda)=\max_{\lambda}\{\min_{\omega,b,\xi_i}L(\boldsymbol{\omega},b,\xi_i,\lambda)\} \quad (10)$$

Three optimal conditions can be derived from equation (9) as follows:

$$\frac{\delta}{\delta\boldsymbol{\omega}}L(\boldsymbol{\omega},b,\xi_i,\lambda)=\boldsymbol{\omega}-\sum_{i=1}^{N}\lambda_i y_i\boldsymbol{x}_i=0 \quad (11)$$

$$\frac{\delta}{\delta b}L(\boldsymbol{\omega},b,\xi_i,\lambda)=\sum_{i=1}^{N}\lambda_i y_i=0 \quad (12)$$

and

$$\frac{\delta}{\delta\xi_i}L(\boldsymbol{\omega},b,\xi_i,\lambda)=Ca_i-\lambda_i-\gamma_i=0 \quad (13)$$

The dual objective function can be obtained by substituting equations (11), (12) and (13) into the right side of the Lagrange function (9). Therefore, the dual problem for the weighted SVM can be formulated as follows:
Maximized:

$$R(\lambda)=\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \quad (14)$$

with constraints defined in SVM and

$$0\le\lambda_i\le a_iC\ ;\qquad i=1,2,..,N \quad (15)$$

It can be seen that by setting $a_i=1$ for all *i*, the weighted support vector machines will be similar to the support vector machines. There is only one free parameter (i.e., *C*) in support vector machines; whereas, in addition to *C*, the number of free parameters in weighted support vector machines is equal to the number of training samples.

It has been observed that the face individuals are highly non-linear because of the variations in facial expression, illumination condition, pose, etc. So, it is necessary to non-linearly map each sample into a high-dimensional feature space using a non-liner function $\varphi:\Re^d\to\Re^D;D>>d$, and then the linear support vector machines can be implemented in high dimensional feature space. A positive definite kernel function *K* is selected *a priori* to perform inner product of vectors in the feature space to avoid explicit mapping $\varphi$ and computational burden in the high-dimensional feature space. The kernel function can be defined as follows:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j) \quad (16)$$

where, $\varphi(\boldsymbol{x}_i)$ is the transformed vector of the person $\boldsymbol{x}_i$ by the non-linear function $\varphi$.

The polynomial and Gaussian radial basis function kernels are two well-known kernel functions:
Polynomial kernel:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)^r \quad (17)$$

Gaussian radial basis function:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right) \quad (18)$$

where, *r* is a positive integer and $\sigma>0$.

In the proposed probabilistic weighted multi-class support vector machines, we used the Gaussian radial basis function as kernel function. Therefore, the dual objective function (14) can be rewritten as follows:
Maximized:

$$F(\lambda)=\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \quad (19)$$

with constraints defined in equations (15).

It can be observed that the objective function to be maximized for the dual problem of the support vector machines and weighted support vector machines is the same. The support vector machines are differing from the weighted support vector machines in that the constraint $0\le\lambda_i\le C$ is replaced with more stringent constraint $0\le\lambda_i\le a_iC$. The constraint optimization for the weighted support vector machines, and computations of the optimum values of the weight vector $\boldsymbol{w}_o$ and bias $b_o$ proceed in the same way as in the case of the support vector machines.

Solving equation (19) with constraints defined in equations (15) determines the optimum Lagrange multipliers $\lambda_{o,i}$. Putting the values of optimum Lagrange multipliers $\lambda_{o,i}$ in equation, the optimum weight vector $\boldsymbol{\omega}_o$ can be obtained.

The Karush-Kuhn-Tucker (KKT) conditions in case of weighted support vector machines can be defined

$$\gamma_i\xi_i=0;\quad i=1,2,...,N \quad (20)$$

By combining equations (13) and (20) the following equation can be formed:

$$(a_iC-\lambda_i)\xi_i=0;\quad i=1,2,...,N \quad (21)$$

From SVM, it can be observed that

$$\xi_i=0\ ;\text{If}\ \lambda_i<a_iC \quad (22)$$

The optimum bias $b_o$is determined by taking any data point $(\boldsymbol{x}_i, y_i)$ in the training set for which $0 < \lambda_{o,i} < a_i C$, and therefore $\xi_i = 0$, and using that data point.

In the proposed probabilistic weighted multi-class support vector machines, we solved the dual objective function using the sequential minimal optimization (SMO) algorithm [20].

# 5 Weighted multi-class support vector machines

The weighted multi-class support vector machines are constructed using a combination of the weighted support vector machines and the decision strategy to decide the class of the input pattern. Each weighted SVM is separately trained. The weighted multi-class support vector machines can be implemented using the one-against-all [1] and one-against-one [21] decision strategies. The one-against-all decision strategy is adopted in the proposed probabilistic weighted multi-class SVM to classify samples, as it requires less amount of memory. This decision strategy is stated as follows:

Let the training set $T = \{\boldsymbol{x}_i, c_j, a_i\}; i \in \{1,2,...N\}; j \in \{1,2,..., M\}$ be the collection of the training sample, its class, and weight, respectively. We designed the weighted SVM for each class by discriminating that class from the rest of (*M*-1) classes. Therefore, in this methodology, we have used *M* number of weighted support vector machines. The set of training samples and their required outputs ($\boldsymbol{x}_i$, $y_i$) are used to design the weighted SVM for class *l*. For a training sample $\boldsymbol{x}_i$, the required output $y_i$ is formulated as follows:

$$y_i = \begin{cases} +1 & if \quad c_j = l \\ -1 & if \quad c_j \neq l \end{cases} \tag{23}$$

The desired output of the *positive* and *negative* samples are $y_i = +1$ and $y_i = -1$, respectively.

The classifier recognizes a test sample by using the *winner-takes-all* decision strategy. Let the test sample $\boldsymbol{x}$ is recognized as class *c*. The output of the classifier is defined as follows:

$$c = \max_{\arg}\{f_l(\boldsymbol{x})\}; \qquad l = 1,2,..., M \tag{24}$$

where, $f_l(\boldsymbol{x})$ is the output of the discriminant function of the weighted SVM constructed for class l.

# 6 Empirical results

We evaluate the performance of the proposed probabilistic weighted multi-class support vector machines on the ***AR face database*** [22], [23], ***CMU PIE face database*** [24], and ***FERET face database*** [25]. Figure 1 (i), (ii), and (iii) displays the face images of a individual from the AR, CMU PIE, FERET face database. The effectiveness of the weighted multi-class support vector machines has also been tested on a synthetic dataset.

The AR face record contains of 26 different frontal subject faces of 126 individual, among them 56 females and 70 males. Individuals are collected in two different sessions divided by two weeks with variation in facial expressions, illumination condition, and occlusion [22, 23].In the CMU PIE face database, there are 41,368 face images of 68 persons (subject) each of 13 different poses, 43 different illumination conditions, and 4 different

(i) (ii) (iii)

Figure 1: Few face images of person from the (i) AR, (ii) CMU PIE (iii) FERET face database.

| Classifier | Recognition rate (%) | |
|---|---|---|
| | 1st Experimental Strategy | 2nd Experimental Strategy |
| **Probabilistic weighted multi-class support vector machines** | **82.50 (38×38)** | **62.50 (36×36)** |
| Multi-class support vector machines | 82.00 (38×38) | 61.91 (36×36) |

Table 1: Comparisons among the probabilistic weighted multi-class support vector machines and the multi-class support vector machines in terms of recognition rates using the performance evaluation over time (first) and performance evaluation with occluded images (second)experimental strategy on the AR face database. In table within the parentheses represent the number of features size.

| Classifier | Avg. recognition rate (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | first experimental strategy | | | | second experimental strategy | | | |
| | *k*=5 | *k*=10 | *k*=15 | *k*=20 | *k*=5 | *k*=10 | *k*=15 | *k*=20 |
| **Probabilistic weighted multi-class support vector machines** | **75.31 (26×26)** | **86.56 (24×24)** | **88.65 (20×20)** | **89.04 (20×20)** | **80.86 (24×24)** | **86.53 (20×20)** | **92.78 (20×20)** | **98.18 (20×20)** |
| Multi-class support vector machines | 75.28 (26×26) | 86.52 (24×24) | 88.59 (20×20) | 88.96 (20×20) | 80.32 (24×24) | 85.86 (20×20) | 91.89 (20×20) | 97.49 (20×20) |

Table 2: Comparisons of the probabilistic weighted multi-class support vector machines and the multi-class support vector machines in terms of average recognition rates for the performance evaluation with pose and expression variations (first) and performance evaluation with illumination variation (second) experimental strategy on the CMU PIE face database. Figures within the parentheses denote the number of features.

| Classifier | Recognition rate (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | FERET Tests September 1996 testing methodology | | | | FRVT 2000 Tests May 2000 testing methodology | | | |
| | fafb | fafc | Dup I | Dup II | P1_probe | P2_probe | P3_probe | P4_probe |
| **Probabilistic weighted multi-class support vector machines** | **98.33 (20×20)** | **97.94 (18×18)** | **89.34 (22×22)** | **83.76 (18×18)** | **68.50 (20×20)** | **49.25 (22×22)** | **28.50 (22×22)** | **22.25 (24×24)** |
| Multi-class support vector machines | 98.16 (20×20) | 96.91 (18×18) | 88.78 (22×22) | 83.33 (18×18) | 67.75 (20×20) | 48.75 (22×22) | 27.75 (22×22) | 21.75 (24×24) |

Table 3: Comparison of performances between the probabilistic weighted multi-class support vector machines and the multi-class support vector machines in terms of recognition rates using FERET Tests September 1996 testing methodology and FRVT 2000 Tests May 2000 testing methodology on the FERET face database. Figures within the parentheses denote the number of features.

expressions. The FERET face database [25] is used to measure the ability of the face recognition system to handle large databases, changes in people's appearance over time, variations in illumination, scale, and pose. Figure 1 (iii) shows example images of a subject from the FERET face database. In this work, experiments are carried out using two standard testing methodology, namely, i) FERET Tests September 1996 testing methodology, and ii) FRVT Tests May 2000 testing methodology. In FERET Tests September 1996 testing methodology, the frontal face images of 1196 subjects are present. The training set contains 1196 face images, one image from each of 1196 distinct subjects. In this testing methodology, there are four test sets, namely, fafb, fafc, Dup I and Dup II. The test sets fafb, fafc, Dup I and Dup II contain 1195, 194, 722 and 234 images, respectively. In FRVT Tests May 2000 testing methodology, the face images of 200 subjects are present. The training set contains 200 frontal images, one image per subject from 200 distinct subjects. In this

testing methodology, there are four test sets, namely, P1_probe, P2_probe, P3_probe and P4_probe.

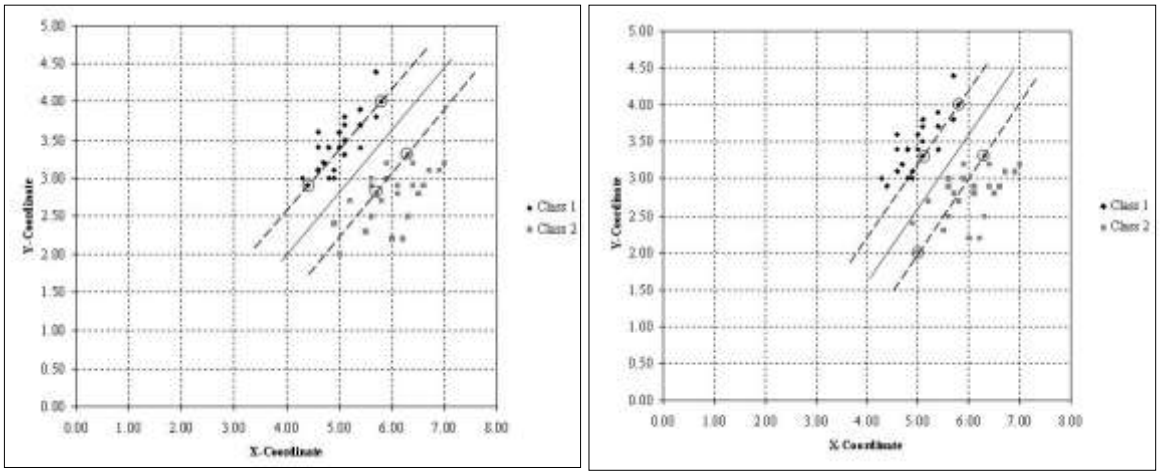
The comparison of performances between the probabilistic weighted multi-class support vector machines and the multi-class support vector machines in terms of recognition rates are illustrated in Table 1, 2, 3 on the AR, CMU-PIE, and FERET face database, respectively. From experimental results, it can be again observed that the performance of the probabilistic weighted multi-class support vector machines is better than the multi-class support vector machines in terms of recognition rate.

In this experiment, a ***synthetic dataset*** *E* containing 2D data from two different classes is randomly generated. In this dataset there are 50 data points, where 25 data points belong to one class and remaining 25 data points belong to another class. Let the dataset *E* can be defined as follows:

$$E = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{50}; \quad \boldsymbol{x}_i \in \Re^2; \quad y_i \in \{+1, -1\} \tag{25}$$

To test the effectiveness of the proposed probabilistic weighted multi-class support vector machines, the data present in the dataset *E* are separately applied on both the multi-class support vector machines as well as on the probabilistic weighted multi-class support vector machines. The optimal separating hyperplane generated by the multi-class support vector machines and the probabilistic weighted multi-class support vector machines are shown in Figures 2(a) and 2(b), respectively.

The encircled data points are support vectors and the distance between the two dotted lines is the margin of separation between two classes in both Figures. The line between these two dotted lines is optimal separating hyperplane. In case of the multi-class support vector machines, 11 data points are present within the margin of separation region, as shown in Figure 2(a). Whereas, in case of the probabilistic weighted multi-class support vector machines, 10 data points are present within the margin of separation region, as shown in Figure 2(b). Therefore, the probabilistic weighted multi-class support vector machines successfully reduces the probability of misclassification, and produces better generalization than that with the multi-class support vector machines.




(a) (b)

Figure 2: Comparative study in terms of the optimal separating hyperplane generation (a): multi-class support vector machines (b): proposed probabilistic weighted multi-class support vector machines, on the dataset *E*.

# 7 Conclusion

In this paper, we present the probabilistic weighted multi-class support vector machines for efficient face recognition. Support vector machines usually used for pattern classification and recognition as well as computer vision domains due to its high generalization ability. However, support vector machines have some limitations because it treats all the training data points of a given class uniformly. As a result, in presence of outliers the training algorithm of the support vector machines can make the decision boundary to be deviated severely from the optimal hyperplane. This limitation of support vector machines can be overcome by the weighted support vector machines where each data point is treated separately according to its weight. In the proposed probabilistic weighted multi-class support vector machines, a reliable weighting model is developed where higher weights are assigned to reliable data points, and lower weights are assigned to outliers. These weights are generated by the probabilistic method; therefore it will take more computing times due to the weight generating algorithm. The training algorithm of the probabilistic weighted support vector machines learns the decision surface according to the relative importance of the training data. The proposed probabilistic weighted multi-class support vector machines have been constructed using a combination of weighted binary support vector machines and one-against-all decision strategy. Several experiments have been carried out on the AR, CMU PIE and FERET face databases using different experimental strategies. The facial features extracted by the G-2DFLD method are separately applied on both the proposed probabilistic multi-class support vector machines as well as on the weighted multi-class support vector machines for training, classification and recognition. The

experimental results show that the performance of the probabilistic weighted multi-class support vector machines is superior to the multi-class support vector machines in terms of recognition rate.

# 8 Acknowledgement

The authors would also like to thank Dr. Sayan Kahali for several discussions which improve the presentation of the paper considerably.

# 9 References

[1] L.J. Cao, K.S. Chau, W.K. Chong, H.P. Lee, and Q.M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine", Neurocomputing, Vol. 55, No. 1-2, pp. 321-336, 2003. https://doi.org/10.1016/S0925-2312(03)00433-8

[2] V.N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[3] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121-167, 1998. https://doi.org/10.1023/A:1009715923555

[4] X. Zhang, "Using class-center vectors to build support vector machines", *Proc. of the IEEE Signal Processing Society Workshop*, pp. 3-11, 1999.

[5] R. Herbrich, and J. Wetson, "Adaptive margin support vector machines for classification", *Proc of the Ninth International Conference on Artificial Neural Networks*, Vol. 2, pp. 880-885, 1999. https://doi.org/10.1049/cp:19991223

[6] Q. Song, W. Hu, and W Xie, "Robust support vector machine with Bullet hole image classification", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 32, No 4, pp. 440-448, 2002. https://doi.org/10.1109/TSMCC.2002.807277

[7] W.J. Hu, and Q. Song, "An accelerated decomposition algorithm for robust support vector machines", *IEEE Transactions on Circuits and Systems II*, Vol. 51, No. 5, pp. 234-240, 2004. https://doi.org/10.1109/TCSII.2004.824044

[8] C. Lin, and S. Wang, "Fuzzy support vector machines", *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 464-471, 2002. https://doi.org/10.1109/72.991432

[9] C. Lin, and S. Wang, "Training algorithms for fuzzy support vector machines with noisy data", *Pattern Recognition Letters*, Vol. 25, No. 2, pp. 1647-1656, 2004. https://doi.org/10.1016/j.patrec.2004.06.009

[10] L.J. Cao, H.P. Lee, and W.K. Chong, "Modified support vector novelty detector using training data with outliers", *Pattern Recognition Letters*, Vol. 24, No. 14, pp. 2479-2487, 2003. https://doi.org/10.1016/S0167-8655(03)00093-X

[11] [11] T. Quan, X. Liu, and Q.Liu, "Weighted least squares support vector machine local region method for nonlinear time series prediction", *Applied Soft Computing*, Vol. 10, No. 2, pp. 562-566, 2010. https://doi.org/10.1016/j.asoc.2009.08.025

[12] J. P. Hwang, S. Park, and E. Kim, "A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function", *Expert Systems with Applications*, Vol. 38, No. 7, pp. 8580-8585, 2011. https://doi.org/10.1016/j.eswa.2011.01.061

[13] L. Yu, "An evolutionary programming based asymmetric weighted least squares support vector machine ensemble learning methodology for software repository mining", *Information Sciences*, Vol. 191, pp. 31-46, 2012. https://doi.org/10.1016/j.ins.2011.09.034

[14] Q. Ye, C. Zhao, S. Gao, and H. Zheng, "Weighted twin support vector machines with local information and its application", *Neural Networks*, Vol. 35, pp. 31-39, 2012. https://doi.org/10.1016/j.neunet.2012.06.010

[15] Y. Shao, W.Chen, J. Zhang, Z. Wang, and N. Deng, "An efficient weighted Lagrangian twin support vector machine for imbalanced data classification", *Pattern Recognition*, Vol. 47, No. 9, pp. 3158-3167, 2014. https://doi.org/10.1016/j.patcog.2014.03.008

[16] P. Xanthopoulos, and T. Razzaghi, "A weighted support vector machine method for control chart pattern recognition", *Computers & Industrial Engineering*, Vol. 70, pp. 134-149, 2014. https://doi.org/10.1016/j.cie.2014.01.014

[17] X. Yang, L. Tan, and L. He, "A robust least squares support vector machine for regression and classification with noise", *Neurocomputing*, Vol. 140, pp. 41-52, 2014. https://doi.org/10.1016/j.neucom.2014.03.037

[18] S. Chowdhury, J.K. Sing, D.K. Basu, and M. Nasipuri, "Face recognition by generalized two-dimensional FLD method and multi-class support vector machines", *Applied Soft Computing*, Vol. 11, No. 7, pp. 4282-4292, 2011. https://doi.org/10.1016/j.asoc.2010.12.002

[19] C. Cortes, and V. Vapnik, "Support-vector network", *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995. https://doi.org/10.1007/BF00994018

[20] J. Platt, "Fast training of support vector machines using sequential minimal optimization", *Advances in Kernel Methods-Support Vector Learning*, MIT Press, Cambridge, pp. 185-208, 1999.

[21] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network", *Neurocomputing*, Vol. 68, pp. 41-50, 1990. https://doi.org/10.1007/978-3-642-76153-9_5

[22] A.M. Martinez, and R. Benavente, "The AR face database", *CVC Technical Report. #24*, June 1998.

[23] A.M. Martinez, and A.C. Kak, "PCA versus LDA", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 228-233, 2001.

https://doi.org/10.1109/34.908974

[24] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database", *Proc. of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-51, 2002.

[25] P.J. Phillips, H. Wechsler, J. Huang, and P.J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms", *Image and Vision Computing*, Vol. 16, No. 5, pp. 295-306, 1998. https://doi.org/10.1016/S0262-8856(97)00070-X

[26] Y. Zhang, Z. Yang, H. Lu, X. Zhou, P. Phillips, Q. Liu, and S. WANG, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation", *Emotion-aware Mobile Computing*, Vol. 4, pp. 8375 – 8385, 2016. https://doi.org/10.1109/ACCESS.2016.2628407

[27] S. Wang, P. Phillips, Z. Dong, Y. Zhang, "Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm", *Neurocomputing*, Vol. 272, pp. 668-676, 2018. https://doi.org/10.1016/j.neucom.2017.08.015

[28] A.A. Aburomman, M.B.I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system", Applied Soft Computing. Vol. 38, pp. 360–372. 2016. https://doi.org/10.1016/j.asoc.2015.10.011

[29] P. Huang, C. Chen, Z. Tang, and Z. Yang, "Feature extraction using local structure preserving discriminant analysis", *Neurocomputing*, Vol. 140, pp. 104-113, 2014. https://doi.org/10.1016/j.neucom.2014.03.031

[30] P. Huang, C. Chen, Z. Tang, and Z. Yang, "Discriminant similarity and variance preserving projection for feature extraction," *Neurocomputing*, Vol. 139, pp. 180-188, 2014. https://doi.org/10.1016/j.neucom.2014.02.047

[31] L. Shi, X. Wang, and Y. Shen, "Research on 3D face recognition method based on LBP and SVM" *Optik*, Vol. 220, pp., 2020. https://doi.org/10.1016/j.ijleo.2020.165157

[32] L. Hu, J. Cui"Digital image recognition based on Fractional-order-PCA-SVM coupling algorithm" *Measurement*, Vol. 145, pp. 150 -159, 2019. https://doi.org/10.1016/j.measurement.2019.02.006

[33] I. Dagher, F. Azar"Improving the SVM gender classification accuracy using clustering and incremental learning", *WileyExpert System*, Vol. 145, pp. 1 -17, 2019. https://doi.org/10.1111/exsy.12372

[34] N.B. Kar, K.S. Babu, A.K. Sangaiah, S Bakshi"Face expression recognition system based on ripplet transform type II and least square SVM"*Multimedia Tools Application*, Vol. 78, pp.4789–4812, 2019. https://doi.org/10.1007/s11042-017-5485-0

# Formal Verification Issues For Component-Based Development

Mehdi Hariati
Computer Science Department, LISCO Laboratory Badji Mokhtar-Annaba University, Annaba, Algeria
E-mail: mehdi.hariati@gmail.com

**Keywords:** component-based development, formal verification, classification.

**Received:** May 6, 2020.

*Component-based development has made a breakthrough in software industry, it offers safer systems and easier to maintain, furthermore, costs and time to market are reduced. However, several issues, such as the correctness of component-based systems, their adaptation or the interactions between their components, require rigorous verification through the use of formal methods and tools. In this paper, we first present an introduction to component-based development; afterward we propose a classification of formal verification issues for component-based systems.*

*Povzetek: V tem članku je predstavljena klasifikacija formalnih metod preverjanja za sisteme, ki temeljijo na komponentah.*

## 1 Introduction

In component-based development [1] the construction of a software system is reduced to an assembly of separately developed software components. This offers as advantages to reduce development costs as well as time to market. Moreover, the quality of the software systems is better, since the latter are built from tested and certified components. In addition, the maintenance and evolution stages of the system are simply a replacement of software components; furthermore, in response to changes in users' requirements or in the environment, component-based systems can also be reconfigured by modifying the links of their architecture.

Nevertheless, the component-based development process should be controlled by the use of formal methods, which allow, at any stage of the lifecycle, verifying important issues; such as the correctness of component-based systems, their adaptation or the interactions between their software components.

This paper is structured as follows. Section 2 presents the basic concepts of component-based development. In Section 3 we show the need for the use of formal methods through a classification of the various verification issues for component-based systems. Section 4 is devoted for related work . Section 5 presents a typical application domain, namely, Web Services. Finally, section 6 concludes this paper.

## 2 Related work

As to the best of our knowledge, this paper is the first presenting a classification of the main issues of formal verifications for the component-based systems, nevertheless, other works deal with the need for the formalization in this domain. In [30], the authors present the need for an abstract approach, the need for formalization for architecture description languages and interface description languages, and the formal languages used for formalization. Compared to our work, the authors invest much more in the study and comparison of the formal languages used in the field of software components, while our work rather focuses on the identification and classification of the problems that may arise during the component based development.

The authors of [31] present briefly an introduction to the component-based development; afterwards the need for formalization in this context is illustrated through a non-trivial example. However, the authors do not offer a detailed classification of potential problems of component-based development.

In [29], a classification of component models is proposed through a comparative study in five dimensions: life cycle, interface specification, interactions, extra-functional properties, and domains. Indeed, this work constitutes a more general classification of component models; the authors introduce the use of formal languages for software components, however, they do not provide a detailed study of formal verification issues for component-based development.

Further, unlike [30] and [31], in order to be more self-contained, basic concepts related to component-based development are provided, this is essential for understanding the formal verification issues.

## 3 Basic concepts of component-based development

In this section we present the basic principles and concepts of component-based development.

### 3.1 Software component

In the literature, there are many definitions of the notion of software component; according to [1], "A software

component is a unit of composition with contractually specified interfaces and explicit context dependencies only. A software component can be deployed independently and is subject to third-party composition". Indeed, a software component interacts with its environment only through its interfaces, since it is designed without any knowledge of its environment; this offers an independence allowing its use in different contexts.

## 3.2 Interfaces and assembly

A software component can have two types of interfaces: on the one hand, the provided interfaces; they represent the services that the component offers, on the other hand, the required interfaces; which are the services that the component needs to accomplish its functions. The assembly of a component-based system is done by linking the provided interfaces with the required interfaces of a selection of software components; however, in order to guarantee a correct assembly of these components, the compatibility of their interfaces should be verified beforehand.

The semantics of an interface is usually specified by its signature. However, the description of an interface only by its signature is insufficient for modeling and verifying the notion of compatibility, indeed, the specification of an interface must also include the definition of the behavior, such as the sequence of service calls between components of the system, or the time constraints, such as the execution time of a service. As we will see in the next sections, the application of formal methods is inescapable for the verification of these issues.

## 3.3 Component models and component frameworks

Others aspects, relating in particular to the definition of the components and their composition are specified by the component model to which the component is assigned. Indeed, the component models define a specific representation, composition modes, interaction styles and others standards dedicated to software components [2]. In addition, component models form the basis for creating *component frameworks*.

Component frameworks establish the physical environmental conditions for the execution and cooperation of components in the system, and they help also to regulate the interactions between components in execution [1].

Component frameworks can only concern physical components, unlike component models, these can be defined for the different levels of abstraction for a component [3], indeed, some component models define a software component as an execution entity, this is the case for Fractal [4] for example, while others component models define a software component as a design entity, as is the case for SOFA [5].

## 3.4 Instance of a software component

Some component models distinguish component types from their instances, allowing the creation and the destruction of component instances at runtime, as is the case for EJB [6] or CCM [7]. Others component models like Wright [8] do not take instantiation into account.

## 3.5 Synchronous communication vs asynchronous communication

Usually, the communication between the software components is done in a synchronous manner, as is the case for Darwin [9] and SOFA. However, in some models such as EJB or CCM, communication can be done by asynchronously sending and receiving messages.

## 3.6 Flat models vs hierarchical models

A set of basic software components can be assembled to give a composite component. In flat component models, this composite component represents the final component-based system, as is the case for EJB or CCM. However, in hierarchical component models, such as SOFA or Fractal, the composite component may in turn be subject to composition with others components, allowing the construction of a component-based system with several hierarchical levels of components. Furthermore, in hierarchical models, we must specify the interfaces to be delegated outside a composite component to be linked to compatible interfaces in the higher hierarchical levels of composition.

## 3.7 Single binding vs multiple binding

Some component models suppose one-to-one linking of interfaces, i.e. single bindings, as in SOFA, others component models allow an interface to be linked to several others interfaces, i.e. multiple bindings, as is the case of EJB and Fractal.

## 3.8 Life cycle of a component-based system

Component-based software systems are developed by selecting and assembling *off-the-shelf components*, instead of being programmed, this makes the lifecycle of a component-based software system different from traditional software system; it mainly comprises the following steps:

1- *Requirements specification*: It concerns collecting, analyzing and specifying the needs of the future users of the system.
2- *Architecture specification*: The architecture of the software specifies the system in terms of abstract components of design and interactions between these components.
3- *Selection and customization of components*: First, the concrete components taken on the shelf are selected according to the software architecture; in a second step, each component must be

personalized before being integrated into the new system.

4- *Integration of the system*: Integration is achieved by establishing mechanisms for communication and coordination of the various components of the final software system.
5- *Test of the system*: Various methods and tools are used to test the component-based system; in fact, it is a question of checking the properties concerning functional aspects as well as those related to the quality of the software.
6- *Deployment*: This is the installation of the software components of the system on one or more computers.
7- *Maintenance and evolution of the system*: After deployment, parts of the component-based system can be modified, due to changes in users' requirements or in the environment.

The concept of software construction by reuse is not new, indeed, the idea was already present in object-oriented programming, it was implemented by the inheritance mechanism; the relatively recent emergence of new technologies has significantly increased the possibilities of building systems and applications from reusable components. Furthermore, building systems based on components or building components for systems in different application areas requires methodologies and processes, including not only development and maintenance aspects, but also those relating to organizational, marketing, legal and other aspects.

## 3.9 Development for reuse and development through reuse

The component-based software engineering process includes two separate but linked processes via a component market. In the following we present each of the two processes:

- *Development for Reuse*: This process consists of an analysis of the application domains in order to develop commercial-off-the-shelf (COTS) components related to these domains. To complete a successful reuse of the software, standards for similar systems must be identified and represented in a form that can be easily exploited to build other systems in the domain. Once created, reusable components will be available in organizations or at the market level as commercial components.
- *Development through reuse*: this is related to the assembly of software systems from the components taken on the shelf.

## 3.10 The objectives of component-based development

The main objectives of component-based development can be summarized as follows:

- *Reuse*: This is the main objective of component-based development. While some software components of a large system are necessarily special purpose components, it is imperative to design and assemble components in order to reuse them in the development of others systems.
- *Independent development of software components*: Large software systems should be able to be assembled from components developed by different people, for this purpose, it is essential to decouple the developers from the components of their users, this is done mainly through the specifications of the behavior of components.
- *Software quality*: A software component or a component-based system should have the desired behavior. Quality assurance technologies for component-based software systems are currently relatively premature, as the characteristics of component-based systems differ from those of conventional systems.
- *Maintainability*: A component-based system should be built in a way that is understandable and easy to evolve.

## 3.11 The contributions of component-based development

The contributions of component-based development can be presented as follows:

- *More efficient management of complexity*: The division of large and complex systems into sub-systems offers greater control over their complexity.
- *Time to market is reduced*: Component-based development consists of assembling existing components, which reduces development time, and therefore accelerates the time to market.
- *Costs are reduced*: While some software components are completely specific to a given application, other software components can be reused and shared with other developers, thereby reducing their costs by damping through a large population.
- *Quality is improved*: Component-based development greatly improves the quality of the systems, since the latter are built from components that are already tested and certified.
- *Easier maintenance and evolution*: The maintenance and evolution of component-based systems is easier, since most of the time they are

reduced to simple additions, deletions or replacement of software components.

# 4 Classification of formal verification issues for component-based systems

Formal approaches are rigorous methods aimed at modeling and analyzing complex systems. The idea of verifying programs is not new; in fact it dates back to the 1960s. Today, formal techniques and tools are widely used in both the academic and the industrial worlds.

In our context, formal methods are essential for component-based development because they enable addressing important verification issues throughout the lifecycle of a component-based system. In the remainder of this section, we will detail these verification issues which we have classified into three levels, namely, at an individual component, during the composition of the components, and finally at the evolution level.

## 4.1 Component level

This level of analysis addresses the verification of an individual component before its composition with the rest of the system; we classified this verification into two types:

- Context-independent verification: it consists of verifying the properties of a component in the isolation, thereby independently of its deployment context; indeed, the issues to be checked can concern the absence of deadlock in its own specification or the coherence of the specification of its temporal constraints.
- Context-dependent verification: In component-based development, components are developed independently of their deployment context; therefore, component correctness can be very difficult to define, as a component may behave correctly in a context but incorrectly in another. Existing approaches remedy this situation in two different ways; some approaches [10, 11] propose to attribute to each component a description of its properties, thereby enabling the user of component to decide if the latter can behave correctly in a given context. Other approaches [12, 13] deliver software components with a set of quality properties that are guaranteed in all contexts satisfying a number of conditions.

## 4.2 Composition level

This level addresses the verification of the composition of the system; we classified this verification into tree main issues:

### 4.2.1 Compatibility of components

The software components constituting a component-based system can be delivered by different sellers; therefore verification of their compatibility is an important issue. Some approaches define compatibility only in terms of signatures of services linking components [14, 7, 15]. However, this description is by no means exhaustive, because it does not include for example, the specification of the services calls sequence of a component, such an aspect is more a matter of behavior. On the other hand, other approaches offer a richer description of compatibility, including description of the behavior [16]. This makes it possible to verify that the composition will not lead to an erroneous interaction between the components of the system.

Some approaches propose to verify compatibility at design time, while others perform checks during execution, thereby detect bad interactions between components dynamically; using a test environment in which the concerned components are duplicated [17].

Moreover, even if the components are not completely incompatible, they can sometimes cooperate correctly by generating appropriate adapters of their interfaces. Some approaches generate adapters for connecting components belonging to different component models [18, 29]; this can be done in a fully automatic manner. Other approaches include adapters for integrating an incompatible functionality of components [19], in which case additional input is required from the user or the monitoring phase to provide information concerning the parts corresponding to the incompatible functionality.

### 4.2.2 Assembly of components

The process of assembling components is mainly twofold: identifying the correct components taken on the shelf, and their connections together, so that the resulting component-based system corresponds to the desired requirements.

Usually, assembly strategies focus on finding the most cost-effective solution with respect to time [19]. The cost function can, for example, evaluate the components in terms of their performance measurements or the minimization of new requirements generated by the added components. The assembly can be selected based on an exhaustive evaluation of all possible alternatives [20], or via an iterative construction of a relatively optimal solution [21].

In this context, formal methods make the problem of assembly of components considerably simpler by simply providing a design of the component based system comprising specifications of a set of components and their connections, the problem being reduced to simply finding the correct component implementations taken on the shelf and formally verifying their compliance with the expected specifications.

### 4.2.3 The global verification

Formal methods are very useful for verifying the global properties of a final component-based system. In this case, formal analysis generally includes:

- Verification of standard coordination errors.

- The absence of deadlock in the system.
- Verification of the different time constraints in the global system.
- The order of execution of a set of services of a components selection in the final system.
- Verification of the number of components that can simultaneously access to the same service.

This verification can be carried out on the whole of the final component-based system or simply on a well-defined part.

Furthermore, in addition to checking properties, formal methods can also help in optimizing component-based systems, namely:

- Detection of inactive components, which can be removed from the system.
- The search for optimal system deployment by placing components in compute nodes based on the density of interaction between them [22].

As with compatibility, some approaches check the properties of a global system at design time, while other approaches allow dynamic verification of the system, in fact, the conformance of the current behavior of the components in execution is verified in parallel with its specification [27], thereby any errors are reported in case of discrepancy.

### 4.3 Evolution level

After the deployment phase, a component-based system can evolve or adapt, in response to changes in users' needs or changes in its environment [23], namely: interoperability with others systems, optimization of computational algorithms, or technical changes.

Formal methods and techniques are very useful for modeling and analyzing the evolution of component-based systems [24]. We have classified this analysis into two types:

- *The dynamic reconfiguration of the architecture*: this mainly includes the change of the links between the system components as well as the creation and destruction of the instances of the components. At this level, formal analysis seeks to verify the coherence of the global system after a dynamic reconfiguration.
- *Substitutability*: one or more components can be replaced with new ones. Generally, approaches addressing this issue define an equivalence relation between the old and the new component, in order to verify that the substitution does not violate the correctness of the global system [25]. However, in some cases, the verification of the equivalence between the two versions of the system is not necessarily strong, because it is only necessary that the new system satisfies a given explicit property, this is considered much more by the approaches that do not aim to guarantee that the behavior remains unchanged, but rather to identify the behavioral differences between several versions of the system [26].

Furthermore, the evolution of a component-based system is usually defined with a set of evolution rules.

## 5 An application domain: Web services

Web services are a typical application domain of component-based development. Indeed, formal methods, used pragmatically, represent a very powerful way to verify several issues, such as the description, composition or evolution of web services.

Regarding the verification of the composition, for instance, the goal is to find the best way to put the services together for the accomplishment of a global task. The composition of web services is called choreography. Nowadays, several languages are dedicated to the description of choreography, for example: WS-CDL (Web Services Choreography Description Language) [32] or WSCI (Web Service Choreography Interface) [33].

Another example of the formal verification for web services is orchestration, this describes the business logic of web services; in fact, it is the description of the control flow of business processes, such as: sequential or parallel execution, etc. WS-BPEL (Web Services Business Process Execution Language) [34] is one of the most widely used languages to describe orchestration.

In this context, formal verification tools perform translations from languages such as: WS-CDL or WS-BPEL, to formalisms, such as: process algebras [8] or timed automata [35], thus allowing the verification of requested properties.

## 6 Conclusion

We presented an overview of the principles and basic concepts of the component-based software development paradigm. Afterwards, through a classification of verification issues for software components, we have shown the need for formal methods and techniques in this context. More generally, for a real integration of formal methods into the component-based development process, frameworks with textual input languages or graphical notations must be provided, and translation algorithms must be implemented; including translations between informal concepts of component-based systems to formalisms, as well as translations of these formalisms to proof or verification tools such as model checking tools.

Further, other issues have yet to be solved. In fact, we have good techniques and tools for formal verifications dedicated to the design phase, such as the UPPAAL model checker [28]; however, these tools cannot be used to do verifications during the execution phase, to control the

behavior of a running system with respect to an expected formal model. On the other hand, it would be practical to design tools that allow direct generation of code from the formal specification of a component-based system.

# 7 Acknowledgement

The authors would like to thank the DGRSDT (General Directorate of Scientific Research and Technological Development) - MESRS (Ministry of Higher Education and Scientific Research), ALGERIA, for the financial support of LISCO Laboratory.

# 8 References

[1] C. Szyperski. Component Software Beyond Object-Oriented Programming. Addison-Wesley, USA, 2. edition, 2002. ISBN 0-201-74572-0.

[2] G. T. Heineman and W. T. Councill. Component Based Software Engineering - Putting the Pieces Together. Addison-Wesley, USA, May 2001. ISBN 0-201-70485-4.

[3] A. Rausch, R. Reussner, and al, editors. The Common Component Modeling Example: Comparing Software Component Models. To appear in LNCS, 2008.

[4] E. Bruneton, T. Coupaye, M. Leclercq, V. Quema, and J.-B. Stefani. The Fractal Component Model and its Support in Java. Software: Practice and Experience, 36(11- 12): 1257-1284, August 2006.

[5] F. Plasil and S. Visnovsky. Behavior Protocols for Software Components. IEEE Transactions on Software Engineering, 28(11): 1056-1076, November 2002.

[6] Sun Microsystems. Enterprise JavaBeans 3.0 Specification, May 2006.

[7] Object Management Group. CORBA Component Model 4.0 Specification. Technical Report formal/06-04-01, Object Management Group, April 2006.

[8] R. J. Allen. A Formal Approach to Software Architecture. PhD thesis, Carnegie Mellon University, School of Computer Science, USA, May 1997.

[9] J. Magee, N. Dulay, S. Eisenbach, and J. Kramer. Specifying Distributed Software Architectures. In Proceedings of the 5th European Software Engineering Conference (ESEC'95), volume 989 of LNCS, pages: 137- 153. Springer-Verlag, September 1995.

[10] B. Meyer. The Grand Challenge of Trusted Components. In Proceedings of the 25th International Conference on Software Engineering (ICSE'03), pages: 660-667. IEEE Computer Society, May 2003.

[11] B. Meyer, C. Mingins, and H. Schmidt. Providing Trusted Components to the Industry. Computer, 31(5): 104-105, May 1998.

[12] G. Xie. Decompositional Verification of Component-based Systems - A Hybrid Approach. In Proceedings of the IEEE International Conference on Automated Software Engineering (ASE'04), pages 414-417. IEEE Computer Society, September 2004.

[13] J. M. Cobleigh, D. Giannakopoulou, and C. S. Pasareanu. Learning Assumptions for Compositional Verification. In Proceedings of the International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'03), volume 2619 of LNCS, pages: 331-346. Springer-Verlag, January 2003.

[14] M. Corporation. COM: Component Object Model Technologies, December 2007. URL http://www.microsoft.com/com/.

[15] N. A. Lynch and M. R. Tuttle. An Introduction to Input/Output Automata. CWI Quarterly, 2(3): 219-246, September 1989.

[16] L. de Alfaro and T. A. Henzinger. Interface-based Design. In Proceedings of the 2004 Marktoberdorf Summer School, pages 1-25. Kluwer, The Netherlands, 2005.

[17] D. Niebuhr and A. Rausch. A concept for dynamic wiring of components: correctness in dynamic adaptive systems. In Proceedings of the ESEC/FSE Conference on Specification and Verification of Component-Based Systems (SAVCBS'07), pages 101-102. ACM Press, September 2007.

[18] O. Galk and T. Bures. Generating Connectors for Heterogeneous Deployment. In Proceedings of the 5th International Workshop on Software Engineering and Middleware (SEM'05), pages 54-61. ACM Press, September 2005.

[19] L. Gesellensetter and S. Glesner. Only the Best Can Make It: Optimal Component Selection. Electronic Notes in Theoretical Computer Science (ENTCS), 176(2): 105-124, May 2007.

[20] N. Barthwal and M. Woodside. E±cient Evaluation of Alternatives for Assembly of Services. In Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'05), pages 1-8. IEEE Computer Society, April 2005.

[21] N. Desnos, S. Vauttier, C. Urtado, and M. Huchard. Software Architecture, volume 4344 of LNCS, chapter Automating the Building of Software Component Architectures, pages 228-235. Springer-Verlag, December 2006.

[22] B. Zimmerova. Component Placement in Distributed Environment w.r.t. Component Interaction. In Proceedings of the Doctoral Workshop on Mathematical and Engineering Methods in Computer Science (MEMICS'06), pages: 260-267. FIT VUT Brno,Czech Republic, October 2006.

[23] P. Waewsawangwong. A Constraint Architectural Description Approach to Self-Organising Component-Based Software Systems. In Proceedings of the International Conference on Software Engineering (ICSE'04), pages: 81-83. IEEE Computer Society,May 2004.

[24] B. Zimmerova and P. Varekova. Reecting Creation and Destruction of Instances in CBSs Modelling and

Verification. In Proceedings of the Doctoral Workshop on Mathematical and Engineering Methods in Computer Science (MEMICS'07), pages: 257-264. Novotny, Brno, Czech Republic, October 2007.

[25] P. Parzek, F. Plasil, and J. Kofro·n. Model Checking of Software Components: Combining Java PathFinder and Behavior Protocol Model Checker. In Proceedings of the Software Engineering Workshop (SEW'06), pages: 133-141. IEEE Computer Society,April 2006.

[26] L. Mariani and M. Pezzµe. A Technique for Verifying Component-Based Software. In Proceedings of the International Workshop on Test and Analysis of Component Based Systems (TACoS'04), volume 116 of ENTCS, pages: 17-30. Elsevier Science Publishers,January 2005.

[27] P. Parzek, F. Plasil, and J. Kofron. Model Checking of Software Components: Combining Java PathFinder and Behavior Protocol Model Checker. In Proceedings of the Software Engineering Workshop (SEW'06), pages: 133-141. IEEE Computer Society,April 2006.

[28] LARSEN, Kim G., PETTERSSON, Paul, et YI, Wang. UPPAAL in a nutshell. International journal on software tools for technology transfer, 1997, vol. 1, no 1-2, p. 134-152.

[29] CRNKOVIC, Ivica, CHAUDRON, Michel, SENTILLES, Séverine, et al. A classification framework for component models. Software Engineering Research and Practice in Sweden, 2007, p. 3.

[30] POIZAT, Pascal, ROYER, Jean-Claude, et SALAÜN, Gwen. Formal methods for component description, coordination and adaptation. Canal et al.[4], 2004, p. 89-100.

[31] MAKOWSKI, Piotr et RAVN, Anders P. Component Based Development-Where is the Place for Formalization?. 2003.

[32] KAVANTZAS, Nickolas, BURDETT, David, RITZINGER, Gregory, et al. Web service choreography description language (wscdl) 1.0. 2004.

[33] ARKIN, Assaf, ASKARY, Sid, FORDIN, Scott, et al. Web service choreography interface (WSCI) 1.0, 2002. URL http://www. w3. org/TR/wsci, 2002.

[34] ARKIN, Assaf, ASKARY, Sid, BLOCH, Ben, et al. Web services business process execution language version 2.0. Working Draft. WS-BPEL TC OASIS, 2005.

[35] ALUR, Rajeev et DILL, David L. A theory of timed automata. Theoretical computer science, 1994, vol. 126, no 2, p. 183-235.

https://doi.org/10.31449/inf.v44i4.3159 Informatica **44** (2020) 477–489 **477**

# Stock Market Decision Support Modeling with Tree-Based Adaboost Ensemble Machine Learning Models

Ernest Kwame Ampomah, Zhiguang Qin, Gabriel Nyame and Francis Effirm Botchey
School of Information & Software Engineering, University of Electronic Science and Technology of China, China
Email: ampomahke@gmail.com, qinzg@uestc.edu.cn, kwakuasane1972@gmail.com, botcheyfrancis@gmail.com

**Keywords:** AdaBoost, machine learning, stock market, features, tree-based ensemble models

**Received:** May 10, 2020

*Forecasting stock market behavior has received tremendous attention from investors and researchers for a very long time due to its potential profitability. Predicting stock market behavior is regarded as one of the extremely challenging applications of time series forecasting. While there is divided opinion on the efficiency of markets, numerous empirical studies which are widely accepted have shown that the stock market is predictable to some extent. Statistical based methods and machine learning models are used to forecast and analyze the stock market. Machine learning (ML) models typically perform better than those of statistical and econometric models. In addition, performance of ensemble ML models is typically superior to those of individual ML models. In this paper, we study and compare the efficiency of tree-based ensemble ML models (namely, Bagging classifier, Random Forest (RF), Extra trees classifier (ET), AdaBoost of Bagging (ADA_of_BAG), AdaBoost of RandomForest (ADA_of_RF), and AdaBoost of ExtraTrees (ADA_of_ET)). Stock data randomly collected from three different stock exchanges were used for the study. Forty technical indicators were computed and used as input features. The data set was spilt into training and test sets. The performance of the models was evaluated with the test set using accuracy, precision, recall, F1-score, specificity and AUC metrics. Kendall W test of concordance was used to rank the performance of the different models. The experimental results indicated that AdaBoost of Bagging (ADA_of_BAG) model was the highest performer among the tree-based ensemble models studied. Also, boosting of the bagging ensemble models improved the performance of the bagging ensemble models.*

*Povzetek: Z Adaboost algoritmi na osnovi dreves je analizirano dogajanje na borzah.*

## 1 Introduction

Forecasting stock market behavior has received tremendous attention from investors, and researchers for a very long time due to its potential profitability (Bacchetta, et al, 2009; Campbell & Hamao, 1992; Granger & Morgenstern, 1970; Lin, et al, 2009; Rajashree & Pradipta, 2016; Weng et, al, 2018). It offers investors the opportunity to be proactive and take decisions which are knowledge-driven in order to gain good returns on their investments with less risk. Predicting stock market behaviour is regarded as one of the extremely challenging applications of time series forecasting. The stock market is affected by factors, such as economic policies, government decrees, political situations, psychology of investors, and so on (Tan, et al, 2007). These factors make the market very dynamic, nonlinear and complex, nonparametric, and chaotic nature (Abu-Mostafa & Atiya, 1996). While there is divided opinion on the efficiency of markets, numerous empirical studies which are widely accepted have shown that the stock market is predictable to some extent (Bollerslev, et al, 2014; Chen, et, al, 2003; Feuerriegel, & Gordon, 2018; Kim, et al, 2011; Phan, et, al, 2015). Statistical based methods and machine learning models are used to forecast and analyze the stock market. The statistical based approaches are not able to predict the stock market very well due the chaotic, noisy and nonlinear in nature of the market. Contrary to statistical approaches, machine learning methods are able deal with the dynamic, chaotic, noisy, and nonlinear data of the stock market and have been widely used for a more accurate forecasting of stock market (Enke & Mehdiyev 2013; Hsu, et al, 2016; Meesad & Rasel, 2013; Thawornwong & Enke 2004; Rather et al. 2015). From the literature, application of machine learning models in stock market prediction can be grouped into a.) application of individual/single machine learning (ML) models (Alkhatib, et al, 2013; Chong et al, 2017; Guresen, et al, 2011; Khansa & Liginlal 2011; Meesad & Rasel 2013; Patel et al. 2015a; Tsai & Hsiao 2010; Wang, et al, 2011; Zhang & Wu 2009). b.) application of ensemble machine learning models. (Araújo, et al, 2015; Booth, et al, 2014; Chen, et al, 2007; Hassan, et al, 2007; Patel et al. 2015b; Rather et al, 2015; Wang, et al, 2012; Wang, et al, 2015). The ensemble models create several individual models to make predictions and then aggregate the outcomes of each individual model to make a final prediction. The performance of ensemble models is better than that of individual models as the ensemble models reduce the generalization error of the predictions. The dominance of ensemble models over individual models has been demonstrated in the field of financial expert systems (Chen et al., 2007; Haung et al, 2008; Tsai et al., 2011). Hence, in this work, we study and compare the effectiveness of tree-based bagging ensemble machine

learning models and the impact of Boosting on the tree-based bagging ensemble models. Specifically, the study compares the effectiveness of the following classifiers: Random forest classifier (RF), Bagging classifier (BAG), and Extra trees classifier (ET), AdaBoost of RandomForest classifier (ADA_of_RF) model, AdaBoost of Bagging classifier (ADA_of_BAG) model and AdaBoost of ExtraTrees classifier (ADA_of_ET) models in forecasting one-day ahead stock price movement.

# 2 Related studies

There have been a number of research studies on forecasting stock market behavior with machine learning algorithms. In this section, we provide a review of some of these studies. Tsai, et al, (2011) studied the performance of ensemble classifiers in analyzing stock returns. They considered the hybrid approaches of majority voting and bagging. They compared the performance homogeneous and heterogeneous ensemble classifiers with those of single baseline classifiers (decision trees, neural networks, and logistic regression). The experimental results indicated that ensemble classifiers outperformed the single classifiers in terms of prediction. In terms of prediction accuracy, there was no significant difference between majority voting and bagging, however, the majority voting had better stock returns than the bagging. Finally, the homogeneous neural networks ensemble classifiers produced the best performance by majority voting when predicting stock returns. Huang et al, (2008) applied wrapper approach to select subset of optimal features from the initial feature set of 23 technical indices and then employed an ensemble voting scheme that combines different classifiers to forecast the trend in Korea and Taiwan stock markets. Experimental outcome shows that the wrapper approach is able to produce better performance than the commonly used features filters, including $\chi 2$ Statistic, Information gain, ReliefF, Symmetrical uncertainty and CFS. In addition, the proposed ensemble voting scheme performed better than the single classifier such as SVM, kth nearest neighbor, back-propagation neural network, decision tree, and logistic regression. Lunga & Marwala, (2006) investigated the predictability of direction of movement of stock market with Learn++ algorithm by predicting the daily movement direction of the Dow Jones. The Learn++ algorithm is derived from the AdaBoost algorithm. The framework was implemented with multi-layer Perceptron (MLP) as a weak Learner. Initially, a weak learning algorithm, which attempts to learn a class concept with a single input Perceptron, is established. The Learn++ algorithm is applied to improve the learning capacity of the weak MLP and introduces the concept of online incremental learning. The proposed framework can adapt as new data are introduced and is able to classify. Balling et al, (2015) compared the performance of ensemble classifier models (Random Forest, AdaBoost and Kernel Factory) against individual classifier models (Neural Networks, Logistic Regression, SVM, and K-Nearest Neighbor). They used data from 5767publicly listed European companies and AUC metric to evaluate the models. The experimental results indicated that Random Forest was the best performer with SVM, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbors and Logistic Regression following in that order. Nayak et al, (2016) made an attempt to predict stock market trend. Two models, one for daily prediction and the other for monthly prediction were built. Three supervised machine learning algorithms namely Decision Boosted Tree, Support Vector Machine, and Logistic Regression were used. With the daily prediction model, historical stock price data were combined with sentiment data. An accuracy of up to 70% were observed using the supervised machine learning algorithms on daily prediction model. It was observed that Decision Boosted Tree performed better than Support Vector Machine and Logistic Regression. The monthly prediction models were used to evaluate the similarity among any two different months trend. The evaluation demonstrated that trend of one month were least correlated with the trend of other months. Khan et al, (2020) employed machine learning algorithms on social media and financial news data to establish the influence of this data on stock market prediction accuracy for ten subsequent days. In order to improve performance and quality of predictions, the authors performed feature selection and spam tweets reduction on the data sets. In addition, experiments to determine stock markets that are difficult to predict and those that are more influence by social media and financial news. A comparison of results of different algorithms to find a consistent classifier was done. Deep learning is used and some classifiers are ensembled. The experimental outcome showed that highest prediction accuracies of 80.53% and 75.16% were attained using social media and financial news, respectively. Also, the results showed that, the New York and Red Hat stock markets are difficult to predict, the New York and IBM stocks are strongly influenced by social media, while London and Microsoft stocks are strongly influenced by financial news. Random forest classifier proved to be consistent and provided the highest accuracy of 83.22% by its ensemble. Nti et al, (2020), conducted a comparative analysis of ensemble machine learning techniques including boosting, bagging, blending and super learners (stacking). The authors build 25 different ensembled regressors and classifiers Using Decision Trees (DT), Support Vector Machine (SVM) and Neural Network (NN). A comparison of their execution times, accuracy, and error metrics over stock-data from Ghana Stock Exchange (GSE), Johannesburg Stock Exchange (JSE), Bombay Stock Exchange (BSE-SENSEX) and New York Stock Exchange (NYSE), from 2012 to 2018 was undertaken. The experimental results showed that stacking and blending ensemble techniques provide higher prediction accuracies (90–100%) and (85.7–100%) respectively, as compared with that of bagging (53–97.78%) and boosting (52.7–96.32%). Also, the root means square error obtained by stacking (0.0001–0.001) and blending (0.002–0.01) provided a better fit of ensemble classifiers and regressors based on these two techniques in market analyses in comparison with bagging (0.01–0.11) and boosting (0.01–0.443). The outcomes suggested that studies in the domain of stock market

direction prediction ought to include ensemble techniques in their sets of algorithms. Vijha et al, (2020) utilized artificial neural network and random forest techniques to predict the next day closing price for five companies which belong to different sectors of operation. The authors generated new variables which are used as inputs to the model from the financial data: Open, High, Low and Close prices of stocks. The evaluation of the models was done using standard RMSE and MAPE.

# 3 Method

The stock data were subjected to (i) data cleaning; to deal with the missing and erroneous values, (ii) data normalization; to ensure that, the machine learning models perform well. Each dataset was split into training and test sets for the purpose of this experiment. The training set was made up of the initial 70% of the data set, and the final 30% of the data set constituted the test set. Each model was trained with the training set and evaluated using the test set.

## 3.1 Data and features

For this research study, we randomly collected ten different stock data from three different stock markets (namely NYSE, NASDAQ, and NSE) through the yahoo finance API. The data from the following companies and indices are used: Apple Inc. ('AAPL'), Abbott Laboratories ('ABT'), Bank of America Corp ('BAC'), Exon mobile corporation ('XOM'), S&P_500 Index, Microsoft Corporation ('MSFT'), Dow Jones Industrial Average Index ('DJIA'), CarMax Inc. ('KMX'), Tata Steel Limited ('TATASTEEL'), and HCL Technologies Ltd ('HCLTECH'). Table 1 provides a description of the

| Data Set | Stock Market | Time Frame | Number of Sample |
|---|---|---|---|
| AAPL | NASDAQ | 2005-01-01 to 2019-12-30 | 3774 |
| ABT | NYSE | 2005-01-01 to 2019-12-30 | 3774 |
| BAC | NYSE | 2005-01-01 to 2019-12-30 | 3774 |
| XOM | NYSE | 2005-01-01 to 2019-12-30 | 3774 |
| S&P_500 | INDEXSP | 2005-01-01 to 2019-12-30 | 3774 |
| MSFT | NASDAQ | 2005-01-01 to 2019-12-30 | 3774 |
| DJIA | INDEXDJX | 2005-01-01 to 2019-12-30 | 3774 |
| KMX | NYSE | 2005-01-01 to 2019-12-30 | 3774 |
| TATASTEEL | NSE | 2005-01-01 to 2019-12-30 | 3279 |
| HCLTECH | NSE | 2005-01-01 to 2019-12-30 | 3477 |

Table 1: Description of the data sets.

data sets used. To ensure generalizability of results, forty (40) technical indicators are computed from the original OHLCV data and used as input features. These technical indicators are selected from four categories of technical indicators which are volume indicators, price transform, overlap studies, and momentum indicators. The details of these technical indicators are provided by table 10-13 in the appendix section.

## 3.2 Feature scaling

The input features have different range of values. Hence, we apply standardization scaling (z-score) to bring all the input features within the same range. The z-score centres values around the mean with a unit standard deviation. The scaling of input features assures that the larger value features do not overwhelm smaller value inputs, and also helps minimize the prediction errors (Kim, 2003).

$$\boldsymbol{z(x) = (x[:, i] - \mu_i)/\sigma_i} \quad (1)$$

Where $\mu_i$ = mean of the $ith$ feature, $\sigma_i$ = standard deviation of the $ith$ feature.

## 3.3 Machine learning algorithms

The study considered and compared the efficacy of Random forest classifier (RF), Bagging classifier (Bag), and Extra trees classifier (ET), AdaBoost of RandomForest (ADA_of_RF) model, AdaBoost of Bagging (ADA_of_BAG) model and AdaBoost of ExtraTrees (ADA_of_ET) in forecasting one-day ahead stock price movement. A discussion of these machine learning (ML) algorithms is presented here.

### 3.3.1 AdaBoost algorithm

AdaBoost is an ensemble/meta-learning approach that builds a strong classifier as a linear combination in an iterative way. In every iteration, it makes a call to a weak learning algorithm (the base learner) which returns a classifier, and gives a weight coefficient to it. AdaBoost tweaks subsequent base learners in favor of those instances misclassified by preceding classifiers. The outcome of the weak learners is aggregated into a weighted sum that represents the final outcome of the boosted classifier. The final output of the boosted classifier is decided by a weighted "vote" of the base classifiers. The smaller the error of the base classifier, the larger is its weight in the final vote (Freund & Schapire, 1996). AdaBoost is sensitive to outliers and noisy data. AdaBoost ML algorithm is given by algorithm 1 below.

### 3.3.2 Decision tree algorithm

Decision tree is a hierarchical tree structure that is used to determine the class label of instances based on a series of if-then rules about the features /attributes of the class. A decision tree consists of nodes (root, internal, and leaf), and branches. The root and internal nodes specify a test condition on a feature, each branch represents one of the possible values of the feature, and each leaf node contains a class label. To classify an instance, we start from the root node and apply the test condition to the instance and follow the branch with the value corresponding to the test outcome. This will take us to either an internal node, for which another test condition is executed, or to a leaf node.

**Input**

Given instances: $\left(x_{1,}y_1\right)...\left(x_{m,}y_m\right); x \to X$ , with labels $y_i \in Y = \{-1, +1\}$

**Initialize:** $D_t\left(i\right) = \frac{1}{m}$ for $i = 1,...,m$.

**for** $t = 1,...,T$ :

1. Call and train a weak learner which returns the weak classifier $h_t : X \to \{-1, 1\}$ with minimum error with respect to distribution $D_t$
2. Compute the error of $h_t$ : $\varepsilon_t = \mathrm{P}_{r \square D_i}\left[h_t\left(x_i\right) \neq y_i\right]$
3. Select $\beta_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
4. Update the distribution $D_{t+1}(i) = \frac{D_t(i)\exp(-\beta_t y_i h_t(x_i))}{Z_t}$

where $Z_t$ a normalization constant is chosen such that $D_{t+1}$ is a distribution

**Output:**

The final hypothesis: $H(x) = sign\left(\sum_{t=1}^{T}\beta_t h_t\left(x\right)\right)$

Algorithm 1: AdaBoost ML algorithm (Freund & Schapire, 1996).

The class label contained in the leaf node is assigned to the instance (Rokach & Maimon, 2008).

### 3.3.3 Bagging algorithm

A Bagging classifier is an ensemble classifier which generates multiple base learners (decision tree) and fits each of these base learners on random subsets of the initial dataset and then combine their individual predictions (through voting or averaging) to produce a final prediction. All the base learners are trained in parallel with the new training sets which are generated by randomly drawing N samples with replacement from the original training dataset – where N is the size of the original training set. The training set for each base learner is independent of the one another. Since the training set for each base learner is generated by resampling initial training data set with replacement, some instances may appear many times while others may not appear. If perturbing the training set can cause significant changes in the models built, then bagging can increase accuracy (Breiman, 1996). Bagging is less sensitivity to outliers and noise, and has a parallel structure for efficient implementations. It is a technique that reduces the variance of an estimated prediction function.

### 3.3.4 Random forest algorithm

Random Forest constructs an ensemble of de-correlated trees and aggregates them to improve upon the robustness and performance of the decision trees (Breiman, 2001). Each tree is trained with a bootstrap sample from the original training data. In addition, a subset of features is selected randomly from the full set of original features to grow the tree at each node. To establish the class label of a new instance, each decision tree delivers a class label for this instance, and random forest then aggregates the class labels predicted and selects the most voted prediction as the label for the new instance. Since RF searches for the best feature among a random subset of features, it leads to a wide diversity that generally produce a better model. RF can handle larger input datasets.

### 3.3.5 Extra trees algorithm

Extra trees algorithm is a tree-based ensemble machine learning algorithm. ET constructs an ensemble of base learners (decision trees) using the classical top-down procedure. The predictions of all the trees are combined to generate the final prediction through majority vote. ET is similar to RF in that it constructs the trees and split nodes with random subsets of features. However, ET differs from RF on two main counts which are (i) ET uses the entire training data to grow the trees (instead of a bootstrap replica). (ii) ET splits nodes by selecting split-points fully at random. The randomization of the cut-point and features together with ensemble averaging reduces variance while the use of the entire original training sample minimizes bias (Geurts, et al, 2006). ET is computationally efficient.

## 3.4 Hyperparameter optimization

Machine learning algorithms have a set of hyperparameters, and these hyperparameters determine how the model is structured. Our aim is to find the right combination of values for these hyperparameters which will ensure that the machine learning models perform at their best. In this work, we set the hyperparameters of the various machine learning algorithms using Bayesian hyperparameter optimization technique (Feurer & Hutter, 2019). Bayesian hyperparameter optimization (BHO) is an iterative technique which has two basic ingredients: a probabilistic surrogate model and an acquisition function to choose the next point to evaluate. In each iteration, the surrogate model is trained on all observations of the target function made so far. The acquisition function then determines the usefulness of various candidate points, trading off exploration and exploitation. It is much cheaper to compute the acquisition function than to evaluate the blackbox function. Therefore, BHO provides an efficient and cheap way to select good hyperparameter for ML models (Bergstra et al, 2011).

| Data Sets | Bag | RF | ET | ADA_of_ BAG | ADA_of_RF | ADA_of_ET |
|---|---|---|---|---|---|---|
| AAPL | 0.9065 | 0.8982 | 0.8861 | 0.9093 | 0.9019 | 0.8824 |
| ABT | 0.8232 | 0.8898 | 0.8852 | 0.8889 | 0.8963 | 0.8843 |
| KMX | 0.9176 | 0.9167 | 0.8889 | 0.9139 | 0.9102 | 0.8722 |
| S&P_500 | 0.9111 | 0.9019 | 0.8852 | 0.9157 | 0.9046 | 0.8926 |
| TATASTEEL | 0.9442 | 0.9378 | 0.9067 | 0.9378 | 0.9356 | 0.9088 |
| HPCL | 0.9203 | 0.9193 | 0.9021 | 0.9294 | 0.9203 | 0.8981 |
| BAC | 0.9028 | 0.8870 | 0.8704 | 0.9065 | 0.8917 | 0.8917 |
| **Mean** | **0.9037** | **0.9072** | **0.8892** | **0.9145** | **0.9087** | **0.8900** |

Table 2: Accuracy Scores of the tree-based ensemble models.

### 3.5 Evaluation metric

The following classical quality evaluation metrics are used to evaluate the performance of the tree-based AdaBoost ensemble ML models: (a) Accuracy, (b) Precision, (c) Recall, (d) F-measure, (e) Specificity, (f) Area under receiver operating characteristics curve (AUC-ROC).

***Accuracy:*** measures the overall number of predictions that the model gets right

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (2)$$

***F1-score:*** provides a harmonic mean of precision and

$$F1\_score = \frac{2\times precision\times recall}{precision+recall} \quad (3)$$

***Specificity:*** assesses how well the classifier is able to identify negative instances.

$$specificity = \frac{tn}{tn+fp} \quad (4)$$

Where $tp$ = true positive, $fp$ = false positive, $tn$ = true negative, and $fn$ = false negative

**ROC curve**: shows the trade-off between true positive to false positive rates.

**AUC**: it tells a model's ability to discriminate between positive and negative instances. The worst AUC is 0.5, and the best AUC is 1.0.

## 4 Results and discussion

The performances of the different tree-based ensemble ML models on the stock data sets are summarized and discussed in this section.

Table 2 displays the accuracy results of the tree-based ensemble models on the various stock data. From this table, the accuracy values of ADA_of_BAG was the best on AAPL, S&P_500, BAC and HPCL stock data sets. Similarly, Bag recorded the highest accuracy values on KMX and TATASTEEL stock data sets. ADA_of_RF obtained the highest accuracy value on the ABT data set.

| Data Sets | Bag | RF | ET | ADA_of_ BAG | ADA_of_RF | ADA_of_ET |
|---|---|---|---|---|---|---|
| AAPL | 0.9130 | 0.9060 | 0.8928 | 0.9160 | 0.9080 | 0.8881 |
| ABT | 0.8210 | 0.8996 | 0.8944 | 0.8936 | 0.9038 | 0.8914 |
| KMX | 0.9190 | 0.9185 | 0.8911 | 0.9154 | 0.9125 | 0.8727 |
| S&P_500 | 0.9184 | 0.9099 | 0.8901 | 0.9214 | 0.9123 | 0.8988 |
| TATASTEEL | 0.9448 | 0.9387 | 0.9085 | 0.9387 | 0.9363 | 0.9091 |
| HPCL | 0.9217 | 0.9205 | 0.9046 | 0.9303 | 0.9209 | 0.9003 |
| BAC | 0.9077 | 0.8939 | 0.8772 | 0.9119 | 0.8992 | 0.8992 |
| **Mean** | **0.9065** | **0.9124** | **0.8941** | **0.9182** | **0.9133** | **0.8942** |

Table 3: F1 Scores of the tree-based ensemble models.

| Data Sets | Bag | RF | ET | ADA_of_ BAG | ADA_of_RF | ADA_of_ET |
|---|---|---|---|---|---|---|
| AAPL | 0.8926 | 0.8748 | 0.8847 | 0.8907 | 0.8966 | 0.8926 |
| ABT | 0.9002 | 0.8543 | 0.8603 | 0.9102 | 0.8822 | 0.8822 |
| KMX | 0.9328 | 0.9271 | 0.9002 | 0.9290 | 0.9156 | 0.9002 |
| S&P_500 | 0.9080 | 0.8978 | 0.9284 | 0.9325 | 0.9018 | 0.9182 |
| TATASTEEL | 0.9457 | 0.9348 | 0.8978 | 0.9348 | 0.9348 | 0.8978 |
| HPCL | 0.9255 | 0.9275 | 0.8986 | 0.9400 | 0.9358 | 0.8986 |
| BAC | 0.8660 | 0.8377 | 0.8302 | 0.8604 | 0.8321 | 0.8321 |
| **Mean** | **0.9101** | **0.8934** | **0.8857** | **0.9139** | **0.8998** | **0.8888** |

Table 4: Specificity Scores of the tree-based ensemble models.

| DataSets | Bag | RF | ET | ADA_of_ BAG | ADA_of_RF | ADA_of_ET |
|---|---|---|---|---|---|---|
| AAPL | 0.9648 | 0.9645 | 0.9562 | 0.9665 | 0.9633 | 0.9469 |
| ABT | 0.9263 | 0.9453 | 0.9564 | 0.9645 | 0.9578 | 0.9516 |
| KMX | 0.9756 | 0.9677 | 0.9558 | 0.9750 | 0.9662 | 0.9453 |
| S&P_500 | 0.9548 | 0.9646 | 0.9623 | 0.9708 | 0.9684 | 0.9608 |
| TATASTEEL | 0.9832 | 0.9814 | 0.9730 | 0.9821 | 0.9809 | 0.9725 |
| HPCL | 0.9766 | 0.9726 | 0.9704 | 0.9792 | 0.9722 | 0.9671 |
| BAC | 0.9644 | 0.9584 | 0.9507 | 0.9716 | 0.9660 | 0.9512 |
| **Mean** | **0.9637** | **0.9649** | **0.9607** | **0.9728** | **0.9678** | **0.9565** |

Table 5: AUC Scores of the tree-based ensemble models.

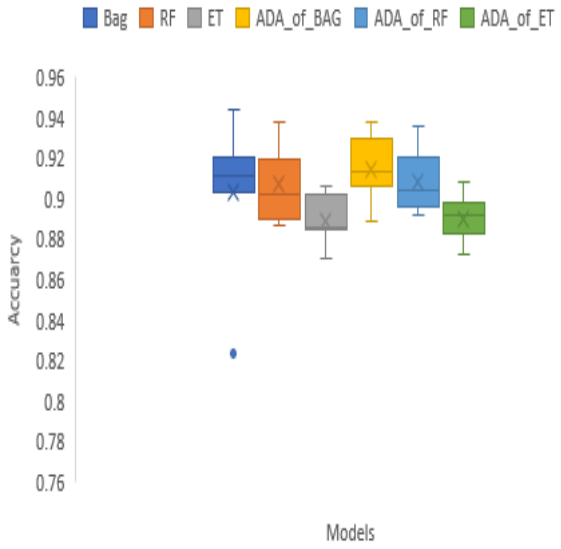



Figure 1: Boxplot of accuracy results of the tree-based ensemble models on the test datasets.

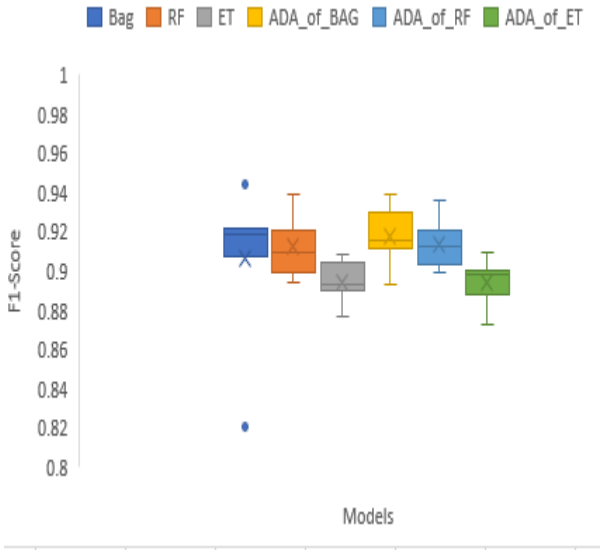



Figure 2: Boxplot of F1-Scores of the tree-based ensemble models on the test datasets.

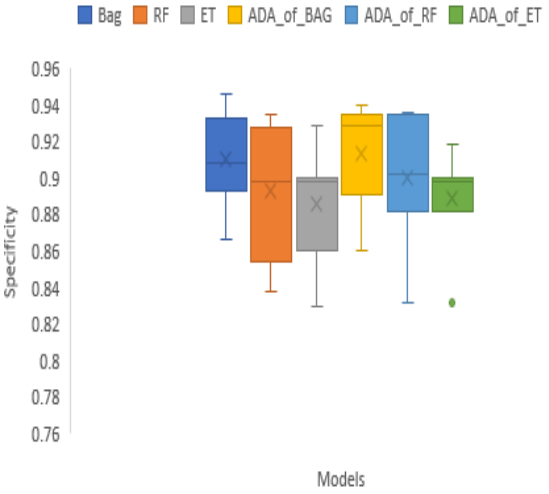



Figure 3: Boxplot of Specificity of the tree-based ensemble models on the test datasets.

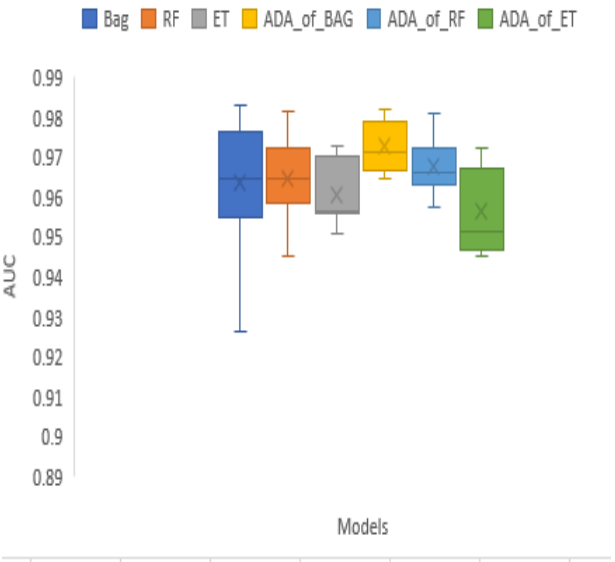



Figure 4: Boxplot of AUC results of the tree-based ensemble models on the test datasets.

Overall, the mean accuracy value of ADA_of_BAG was the best among all the tree-based ensemble algorithms. Boosting of the bagging algorithms (ADA_of_ BAG, ADA_of_RF and ADA_of_ET) improved the mean accuracy values of their respective bagging algorithms (Bag, RF and ET). Figure 1 presents the box plot of the accuracy values of the various models.

Table 3 presents the F1-Scores of the tree-based ensemble models on the various stock data. ADA_of_BAG obtained the highest F1-Score on AAPL, S&P_500, BAC and HPCL stock data sets. Also, Bag recorded the highest accuracy values on KMX and TATASTEEL stock data sets. ADA_of_RF achieved the best F1-Score on the ABT stock data set. In general, the mean F1-value of ADA_of_BAG was the best among all the tree-based ensemble algorithms. In addition, boosting of the bagging algorithms (ADA_of_ BAG, ADA_of_RF and ADA_of_ET) improved the mean F1 values of their respective base bagging algorithms (Bag, RF and ET). Figure 2 presents the box plot of the F1-Scores of the various models.

Table 4 shows the specificity results of the tree-based ensemble models on the various stock data. ADA_of_BAG had the highest specificity on ABT, S&P_500 and HPCL stock data sets. Also, Bag obtained the highest specificity on KMX, TATASTEEL and BAC stock data sets. ADA_of_RF achieved the highest specificity on the ABT stock data set. The mean specificity value of ADA_of_BAG was the best among all the tree-based ensemble algorithms. Moreover, boosting of the bagging algorithms (ADA_of_ BAG, ADA_of_RF and ADA_of_ET) improved the mean specificity results of their respective base bagging algorithms (Bag, RF and ET). Figure 3 presents the box plot of the specificity results of the various models.

Table 5 presents the AUC results of the tree-based ensemble models on the various stock data. ADA_of_BAG performed better than the other models on AAPL, ABT, S&P_500, BAC and HPCL stock data sets.

Similarly, the performance of Bag was higher than the other models on KMX and TATASTEEL stock data sets. In general, the mean AUC of ADA_of_BAG was the best among all the tree-based ensemble algorithms. In addition, boosting of the bagging algorithms Bag and RF (ADA_of_ BAG and ADA_of_RF) recorded a better mean AUC value than their respective base bagging algorithms (Bag and RF). Figure 4 shows the box plot of the AUC results of the various models.

Figure 5-11 shows the ROC curves of all the tree-based ensemble models considered in this study on the AAPL, ABT, KMX, S&P_500, TATASTEEL, HPCL and BAC stock data sets respectively.

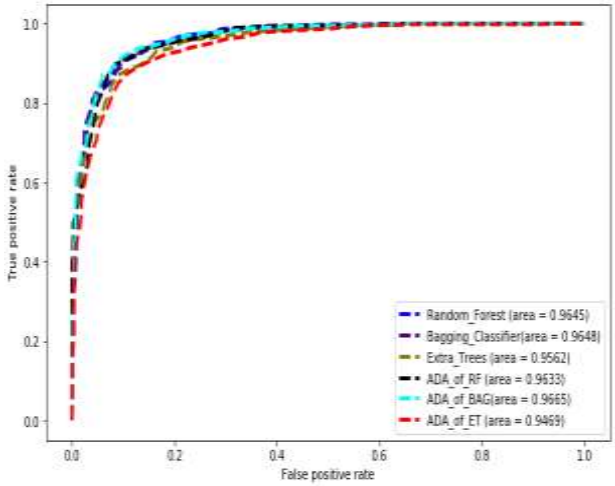



Figure 5: ROC curve of the tree-based ensemble models on AAPL stock data set.

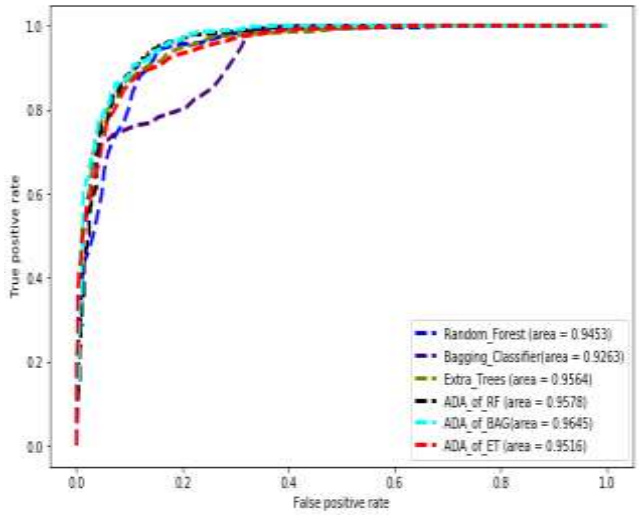



Figure 6: ROC curve of the tree-based ensemble models on ABT stock data set.

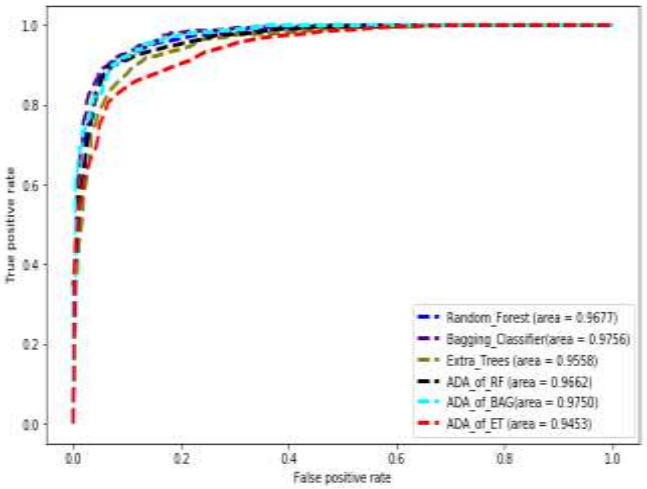



Figure 7: ROC curve of the tree-based ensemble models on KMX stock data set.

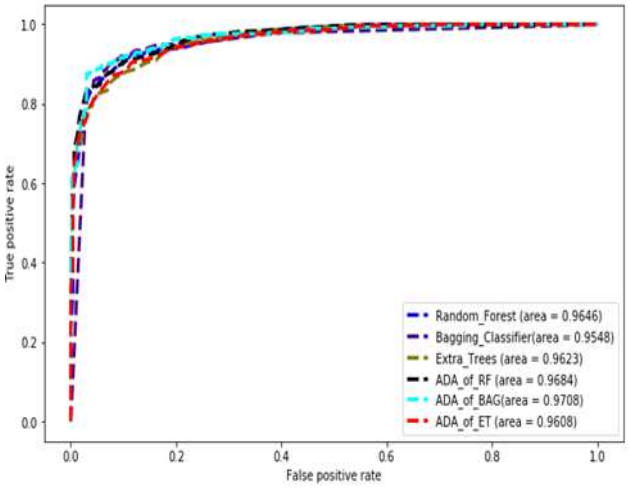



Figure 8: ROC curve of the tree-based ensemble models on S&P_500 stock data set.

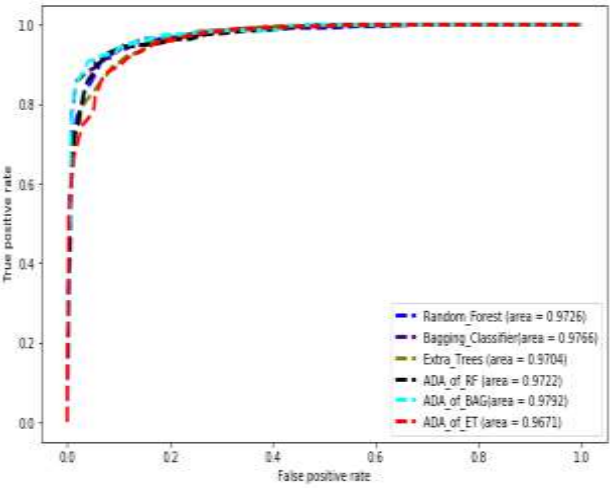



Figure 9: ROC curve of the tree-based ensemble models on HPCL stock data set.

| **Metric** | **W** | $\chi^2$ | **p** | **Ranks** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 0.61 | 21.29 | 0.00 | Technique | Bag | RF | ET | ADA_of _BAG | ADA_ of_RF | ADA_of _ET | |
| | | | | Mean Rank | 4.64 | 3.64 | 1.71 | **5.21** | 4.00 | 1.79 | |

Table 6: Kendall's coefficient of concordance ranks of tree-based ensemble models using accuracy metric.

| **Metric** | **W** | $\chi^2$ | **p** | **Ranks** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| F1-Score | 0.56 | 19.57 | 0.00 | Technique | Bag | RF | ET | ADA_of _BAG | ADA_ of_RF | ADA_of _ET |
| | | | | Mean Rank | 4.71 | 3.64 | 1.86 | **5.07** | 3.93 | 1.79 |

Table 7: Kendall's coefficient of concordance ranks of tree-based ensemble models using F1-score metric.

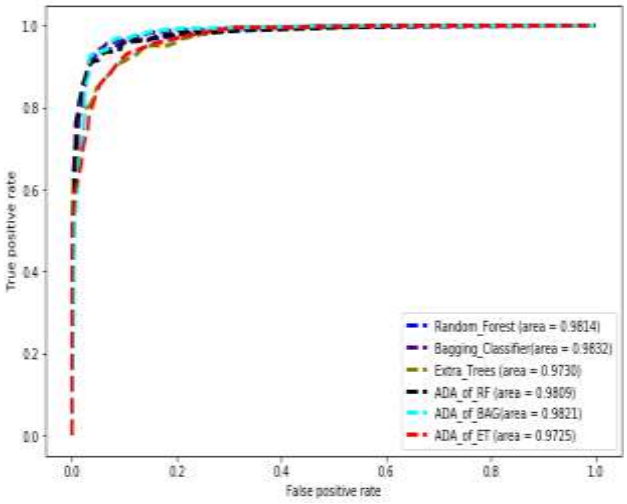



Figure 10: ROC curve of the tree-based ensemble models on TATASTEEL stock data set.

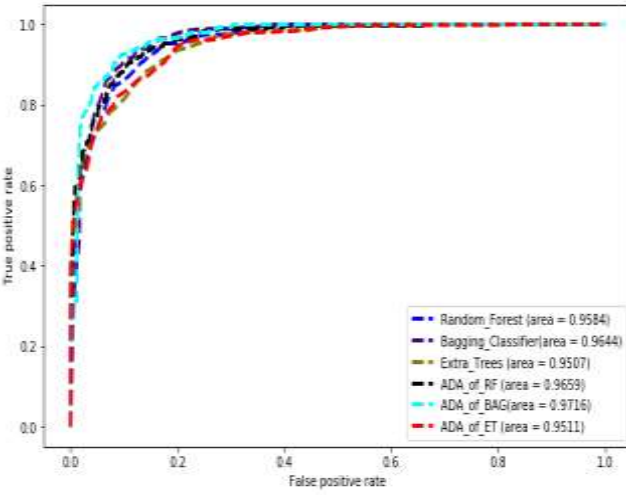



Figure 11: ROC curve of the tree-based ensemble models on S&P_500 stock data set.

The Kendall's coefficient of concordance (W) is applied to rank the efficiency of the different tree-based AdaBoost ensemble models. This test is a measure that applies ranks to establish an agreement among raters (Kendall & Babington, 1939). It determines the agreement among diverse raters who are evaluating a given set of n objects. Depending on the area where it is being applied, the raters can be variables, characters, and so on. The raters are the different data sets in this articleKendall's coefficient of concordance has been applied in many researches including Kendall's Coefficient of Concordance for Sociometric Rankings with Self Excluded by Gordon et al, (1971), Use of Kendall's coefficient of concordance to assess agreement among observers of very high-resolution imagery by Gearhart et al, (2013), Measuring and testing interdependence among random vectors based on Spearman's ρ and Kendall's τ by Zhang & Wang, (2020), In this study a cut-off value of 0.05 for the significance level (p-value) is used. The Kendall's coefficient is considered to be significant and having the capability of giving an overall ranking when $p<0.05$. At $p = 0.05$, the critical value of chi-square ($\chi^2$) for five (5) degrees of freedom is 11.07. The degrees of freedom equal the total number of ML algorithms (which is six) minus one. The results of Kendall's coefficient of concordance are given by tables 6-9 below using accuracy, precision, recall, F1-score, specificity, and AUC respectively.

Table 6 shows that Kendall's coefficient using the accuracy metric is significant ($p<0.05$, $\boldsymbol{\chi^2}>11.07$) and that the performance of ***ADA_of_BAG*** model is the best among the ensemble methods. The overall ranking is ***ADA_of_BAG*** **>Bag >** ***ADA_of_RF*** **> RF >** ***ADA_of_ET*** **>*ET*.**

Table 7 presents that Kendall's coefficient using the F1-Score metric is significant ($p<0.05$, $\boldsymbol{\chi^2}>11.07$) and the performance of ***ADA_of_BAG*** model is the best among the ML ensemble models. The overall ranking is **ADA_of_BAG >Bag > ADA_of_RF > RF > ET > ADA_of_ET.**

Table 8 demonstrates that Kendall's coefficient using the specificity metric is significant ($p>0.05$, $\boldsymbol{\chi^2}<11.07$), and ***ADA_of_BAG*** had the highest rank. The overall ranking is ***ADA_of_BAG*** **>Bag >*ADA_of_RF* > RF = *ADA_of_ET* > *ET*.**

Table 9 demonstrates that Kendall's coefficient using the AUC metric is significant ($p<0.05$, $\boldsymbol{\chi^2}>11.07$) and the performance of ***ADA_of_BAG*** model has the best rank

| **Metric** | **W** | $\chi^2$ | **p** | **Ranks** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Specificity** | 0.41 | 14.45 | 0.01 | Technique | Bag | RF | ET | ADA_of_ BAG | ADA_of _RF | ADA_ of_ET |
| | | | | Mean Rank | 4.79 | 2.71 | 2.07 | **5.00** | 3.71 | 2.71 |

Table 8: Kendall's coefficient of concordance ranks of tree-based ensemble models using specificity metric.

| **Metric** | **W** | $\chi^2$ | **p** | **Ranks** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **AUC** | 0.600 | 20.95 | 0.00 | Technique | Bag | RF | ET | ADA_of_ BAG | ADA_ of_RF | ADA_ of_ET |
| | | | | Mean Rank | 4.00 | 3.57 | 2.29 | **5.71** | 3.86 | 1.58 |

Table 9: Kendall's coefficient of concordance ranks of tree-based ensemble models using AUC metric.

among the tree-based AdaBoost ML ensemble models. The overall ranking is **ADA_of_BAG>Bag >ADA_of_RF > RF > ET > ADA_of_ET**

# 5 Conclusion

This study compares the efficacy of tree-based of bagging ensemble machine learning models and boosting of tree-based bagging machine learning models in forecasting movement direction of stock prices. Seven randomly collected stock data from three different stock exchanges were used. The data sets were split into training and test sets. The performance of the models was evaluated using accuracy, F1-score, specificity, and AUC metrics on the test data set. Kendall W test of concordance was used to ranked the performance of the different models. The results indicated that boosting of tree-based bagging ensemble models, improves the performance of the bagging models. Overall, the performance of ADA_of_BAG model was superior to the remaining models used in the study. The limitation of this study is that it only considered bagging models and boosting of bagging models. Hence, future study will investigate boosting models and bagging of boosting models in predicting stock price behaviour.

# 6 Acknowledgement

This work was supported by the NSFC-Guangdong Joint Fund (Grant No. U1401257), National Natural Science Foundation of China (Grant Nos. 61300090, 61133016, and 61272527), science and technology plan projects in Sichuan Province (Grant No. 2014JY0172) and the opening project of Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology (Grant No. 2013A061401003).

# 7 References

[1] Abu-Mostafa, Y. S., & Atiya, A. F. Introduction to financial forecasting. Applied Intelligence, 6(3), 205–213, 1996. https://doi.org/10.1007/bf00126626

[2] Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. Stock price predic- tion using k-nearest neighbor (knn) algorithm. International Journal of Business, Humanities and Technology, 3 (3), 32–44, 2013.

[3] Ampomah, E., K., Qin Z., & Nyame, G. Evaluation of Tree-based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement, information, 11, 332, 2020. https://doi:10.3390/info11060332

[4] Araújo, R. d. A., Oliveira, A. L., & Meira, S. A hybrid model for high-frequency stock market forecasting. Expert Systems with Applications, 42 (8), 4081–4096, 2015. https://doi.org/10.1016/j.eswa.2015.01.004.

[5] Bacchetta, P., Mertens, E., & Van Wincoop, E. Predictability in financial mar- kets: What do survey expectations tell us? Journal of International Money and Finance, 28 (3), 406–426, 2009. https://doi.org/10.1016/j.jimonfin.2008.09.001

[6] Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42 (20), 7046–7056, 2015. https://doi.org/10.1016/j.eswa.2015.05.013

[7] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24, 2546-2554, 2011.

[8] Bollerslev, T., Marrone, J., Xu, L., & Zhou, H. Stock return predictability and variance risk premia: Statistical inference and international evidence. Journal of Financial and Quantitative Analysis, 49 (03), 633–661, 2014. https://doi.org/10.1017/s0022109014000453

[9] Booth, A., Gerding, E., & Mcgroarty, F. Automated trading with performance weighted random forests and seasonality. Expert Systems with Applications, 41(8), 3651–3661, 2014. https://doi.org/10.1016/j.eswa.2013.12.009

[10] Breiman, L. Bagging predictors. Mach Learn 24, 123–140, 1996. https://doi.org/10.1007/bf00058655

[11] Breiman, L. Random forests. Machine learning, 45(1), 5-32, 2001.

[12] Campbell, J. Y., & Hamao, Y. Predictable stock returns in the united states and japan: A study of long-term capital market integration. The Journal of Finance, 47 (1), 43–69, 1992. https://doi.org/10.1111/j.1540-6261.1992.tb03978.x

[13] Chen, A.-S., Leung, M. T., & Daouk, H. Application of neural networks to an emerging financial market: Forecasting and trading the taiwan stock index.

Computers & Operations Research, 30 (6), 901–923, 2003. https://doi.org/10.1016/s0305-0548(02)00037-0

[14] Chen, Y., Yang, B., & Abraham, A. Flexible neural trees ensemble for stock index modeling. Neurocomputing, 70 (4), 697–703, 2007. https://doi.org/10.1016/j.neucom.2006.10.005

[15] Chong, E., Han, C., & Park, F. C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187–205, 2017. https://doi.org/10.1016/j.eswa.2017.04.030.

[16] Enke, D., & Mehdiyev, N. Stock market prediction using a combination of stepwise regression analysis, differential evolution-based fuzzy clustering, and a fuzzy inference neural network. Intelligent Automation & Soft Computing, 19 (4), 636–648, 2013. https://doi.org/10.1080/10798587.2013.839287

[17] Feurer M., Hutter F. Hyperparameter Optimization. In: Hutter F., Kotthoff L., Vanschoren J. (eds) Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham, 2019.

[18] Feuerriegel, S., & Gordon, J. Long-term stock index forecasting based on text mining of regulatory disclosures. Decision Support Systems, 112: 88–97, 2018. https://doi.org/10.1016/j.dss.2018.06.008

[19] Freund, Y., & Schapire, R. Experiments with a new boosting algorithm. In machine learning. proceedings of the thirteenth international conference (ICML '96). 148–156. Bari, Italy, 1996.

[20] Gearhart, A., Booth, D. T., Sedivec, K. & Schauer, C. Use of Kendall's coefficient of concordance to assess agreement among observers of very high-resolution imagery, Geocarto International, 28(6), 517-526, 2013. https://doi:10.1080/10106049.2012.725775.

[21] Geurts P., Ernst, D., Wehenkel L. Extremely randomized trees, Mach Learn, 63: 3–42, 2006. https://doi.org/10.1007/s10994-006-6226-1

[22] Ghorbani, M., & Chong E., K., P. (2020), Stock price prediction using principal components, PLoS One, 15(3): e0230124. https://doi: 10.1371/journal.pone.0230124.

[23] Gordon H. L., & Richard G. J. Kendall's Coefficient of Concordance for Sociometric Rankings with Self Excluded, Sociometry, 34(4), 496-503, 1971. https://doi.org/10.2307/2786195

[24] Guresen, E., Kayakutlu, G., & Daim, T. U. Using artificial neural network models in stock market index prediction. Expert Systems with Applications, 38 (8), 10389–10397, 2011. https://doi.org/10.1016/j.eswa.2011.02.068

[25] Granger, C. W. J., & Morgenstern, O. Predictability of stock market prices: 1. DC Heath Lexington, Mass. 1970.

[26] Hassan, M. R., Nath, B., & Kirley, M. A fusion model of hmm, ann and ga for stock market forecasting. Expert Systems with Applications, 33 (1), 171–180, 2007. https://doi.org/10.1016/j.eswa.2006.04.007

[27] Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., & Johnson, J. E. Bridging the di- vide in financial market forecasting: Machine learners vs. financial economists. Expert Systems with Applications, 61, 215–234, 2016. https://doi.org/10.1016/j.eswa.2016.05.033

[28] Huang, C.-J., Yang, D.-X., & Chuang, Y.-T. Application of wrapper approach and composite classifier to the stock trend prediction. Expert Systems with Applications, 34(4), 2870–2878, 2008. https://doi.org/10.1016/j.eswa.2007.05.035

[29] Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing, 1-24, 2020. https://doi.org/10.1007/s12652-020-01839-w

[30] Khansa, L., & Liginlal, D. Predicting stock market returns from malicious attacks: A comparative analysis of vector autoregression and time-delayed neural networks. Decision Support Systems, 51 (4), 745–759, 2011. https://doi.org/10.1016/j.dss.2011.01.010

[31] Kim, J. H., Shamsuddin, A., & Lim, K. P. Stock return predictability and the adaptive markets hypothesis: Evidence from century-long us data. Journal of Empirical Finance, 18 (5), 868–879, 2011. https://doi.org/10.1016/j.jempfin.2011.08.002

[32] Meesad, P., & Rasel, R. I. (2013). Predicting stock market price using support vector regression. In Informatics, electronics & vision (iciev), 2013 international confer- ence on (pp. 1–6). IEEE. https://doi.org/10.1109/iciev.2013.6572570

[33] Nayak, A., Pai M., M., M., & Pai R., M. Prediction Models for Indian Stock Market. Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016). Procedia Computer Science 89 441 – 449, 2016. https://doi.org/10.1016/j.procs.2016.06.096

[34] Nti, I. K., Adekoya, A. F., & Weyori, B. A. A comprehensive evaluation of ensemble learning for stock-market prediction. Journal of Big Data, 7(1), 1-40, 2020. https://doi.org/10.1186/s40537-020-00299-5

[35] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42 (1), 259–268, 2015a. https://doi.org/10.1016/j.eswa.2014.07.040

[36] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42 (4), 2162–2172, 2015b. https://doi.org/10.1016/j.eswa.2014.10.031

[37] Phan, D. H. B., Sharma, S. S., & Narayan, P. K. Stock return forecasting: Some new evidence. International Review of Financial Analysis, 40, 38–51, 2015.

https://doi.org/10.1016/j.irfa.2015.05.002
[38] Rajashree D., & Pradipta K., D. A hybrid stock trading framework integrating technical analysis with machine learning techniques. The Journal of Finance and Data Science, 2: 42-57, 2016. https://doi.org/10.1016/j.jfds.2016.03.002
[39] Rather, A. M., Agarwal, A., & Sastry, V. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications, 42 (6), 3234–3241, 2015. https://doi.org/10.1016/j.eswa.2014.12.003
[40] Rokach, L., & Maimon, O. Z. Data mining with decision trees: theory and applications (Vol. 69). World scientific, 2008.
[41] Tan, T. Z., Quek, C., & See, Ng. G. Biological brain-inspired genetic complementary learning for stock market and bank failure prediction. Computational Intelligence, 23(2), 236–261, 2007. https://doi.org/10.1111/j.1467-8640.2007.00303.x
[42] Thawornwong, S., & Enke, D. The adaptive selection of financial and economic variables for use with artificial neural networks. Neurocomputing, 56, 205–232, 2004. https://doi.org/10.1016/j.neucom.2003.05.001
[43] Tsai, C.-F., Lin, Y.-C., Yen, D. C., & Chen, Y.-M. Predicting stock returns by classifier ensembles. Applied Soft Computing, 11 (2), 2452–2459, 2011. https://doi.org/10.1016/j.asoc.2010.10.001
Tsai, C.-F., & Hsiao, Y.-C. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50 (1), 258–269, 2010. https://doi.org/10.1016/j.dss.2010.08.028
[44] Vijha M., Chandolab D., Tikkiwalb V. A., & Kumarc A. Stock Closing Price Prediction using Machine Learning Techniques, Procedia Computer Science Vol. 167, 599-606, 2020. https://doi.org/10.1016/j.procs.2020.03.326
[45] Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., & Guo, S.-P. Stock index forecasting based on a hybrid model. Omega, 40 (6), 758–766, 2012.
[46] https://doi.org/10.1016/j.omega.2011.07.008
[47] Wang, J.-Z., Wang, J.-J., Zhang, Z.-G., & Guo, S.-P. Forecasting stock indices with back propagation neural network. Expert Systems with Applications, 38(11), 14346–14355, 2011. https://doi.org/10.1016/j.eswa.2011.04.222
[48] Wang, L., Zeng, Y., & Chen, T. Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. Expert Systems with Applications, 42 (2), 855–863, 2015. https://doi.org/10.1016/j.eswa.2014.08.018
[49] Weng, B., Lu L., Wang, X., Megahed, F., M., Martinez, W. Predicting short-term stock prices using ensemble methods and online data sources. Expert Systems with Applications, 112: 258–273, 2018. https://doi.org/10.1016/j.eswa.2018.06.016
[50] Zhang, L., Lu, D. & Wang, X. Measuring and testing interdependence among random vectors based on Spearman's ρ and Kendall's τ., Comput Stat, 2020. https://doi.org/10.1007/s00180-020-00973-5
[51] Zhang, Y., & Wu, L. Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. Expert Systems with Applications, 36 (5), 8849–8854, 2009. https://doi.org/10.1016/j.eswa.2008.11.028
[52] Zhong, X., Enke, D. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5, 24, 2019. https://doi.org/10.1186/s40854-019-0138-0

# 8 Appendix

| Volume Indicator | Description |
|---|---|
| Chaikin A/D Line (ADL) | Estimates the Advance/Decline of the market. |
| Chaikin A/D Oscillator (ADOSC) | Indicator of another indicator. It is created through application of MACD to the Chaikin A/D Line |
| On Balance Volume (OBV) | Uses volume flow to forecast changes in price of stock |

Table 10: Description of Volume Indicators used in the study.

| Overlap Studies Indicators | Description |
|---|---|
| Bollinger Bands (BBANDS) | Describes the different highs and lows of a financial instrument in a particular duration. |
| Weighted Moving Average (WMA) | Moving average that assign a greater weight to more recent data points than past data points |
| Exponential Moving Average (EMA) | Weighted moving average that puts greater weight and importance on current data points, however, the rate of decrease between a price and its preceding price is not consistent. |
| Double Exponential Moving Average (DEMA) | It is based on EMA and attempts to provide a smoothed average with less lag than EMA. |
| Kaufman Adaptive Moving Average (KAMA) | Moving average designed to be responsive to market trends and volatility. |
| MESA Adaptive Moving Average (MAMA) | Adjusts to movement in price based on the rate of change of phase as determined by the Hilbert transform discriminator. |
| Midpoint Price over period (MIDPRICE) | Average of the highest close minus lowest close within the look back period |
| Parabolic SAR (SAR) | Heights potential reversals in the direction of market price of securities. |
| Simple Moving Average (SMA) | Arithmetic moving average computed by averaging prices over a given time period. |
| Triple Exponential Moving Average (T3) | It is a triple smoothed combination of the DEMA and EMA |
| Triple Exponential Moving Average (TEMA) | An indicator used for smoothing price fluctuations and filtering out volatility. Provides a moving average having less lag than the classical exponential moving average. |
| Triangular Moving Average (TRIMA) | Moving average that is double smoothed (averaged twice) |

Table 11: Description of Overlap Studies Indicators used in the study.

| **Momentum Indicators** | **Description** |
|---|---|
| Average Directional Movement Index (ADX) | Measures how strong or weak (strength of) a trend is over time |
| Average Directional Movement Index Rating (ADXR) | Estimates momentum change in ADX. |
| Absolute Price Oscillator (APO) | Computes the differences between two moving averages |
| Aroon | Used to find changes in trends in the price of an asset |
| Aroon Oscillator (AROONOSC) | Used to estimate the strength of a trend |
| Balance of Power (BOP) | Measures the strength of buyers and sellers in moving stock prices to the extremes |
| Commodity Channel Index (CCI) | Determine the price level now relative to an average price level over a period of time |
| Chande Momentum Oscillator (CMO) | Estimated by computing the difference between the sum of recent gains and the sum of recent losses |
| Directional Movement Index (DMI) | Indicate the direction of movement of the price of an asset |
| Moving Average Convergence /Divergence (MACD) | Uses moving averages to estimate the momentum of a security asset |
| Money Flow Index (MFI) | Utilize price and volume to identify buying and selling pressures |
| Minus Directional Indicator (MINUS_DI) | Component of ADX and it is used to identify presence of downtrend. |
| Momentum (MOM) | Measurement of price changes of a financial instrument over a period of time |
| Plus Directional Indicator (PLUS_DI) | Component of ADX and it is used to identify presence of uptrend. |
| Log Return | The log return for a period of time is the addition of the log returns of partitions of that period of time. It makes the assumption that return**s** are compounded continuously rather than across sub-periods |
| Percentage Price Oscillator (PPO) | Computes the difference between two moving averages as a percentage of the bigger moving average |
| Rate of change (ROC) | Measure of percentage change between the current price with respect to a at closing price n periods ago. |
| Relative Strength Index (RSI) | Determines the strength of current price in relation to preceding price |
| Stochastic (STOCH) | Measures momentum by comparing closing of a security with earlier trading range over a specific period of time |
| Stochastic Relative Strength Index (STOCHRSI) | Used to estimate whether a security is overbought or oversold. It measures RSI over its own high/low range over a specified period. |
| Ultimate Oscillator (ULTOSC) | Estimates the price momentum of a security asset across different time frames. |
| Williams' %R (WILLR) | Indicates the position of the last closing price relative to the highest and lowest price over a time period. |

Table 12: Description of Momentum Indicators used in the study.

| **Price Transform Indicator** | **Description** |
|---|---|
| Median Price (MEDPRICE) | Measures the mid-point of each day's high and low |
| Typical Price (TYPPRICE) | Measures the average of each day's price. |
| Weighted Close Price (WCLPRICE) | Average of each day's price with extra weight given to the closing price. |

Table 13: Description of Price Transform Indicators used in the study.

# Face Recognition Based on Deep Learning Under the Background of Big Data

Hongbiao Ni
Department of Information Engineering, Jilin Police College, Changchun 130117, Jilin, China
E-mail: nhbhongb@yeah.net

**Keywords:** big data, deep learning, face recognition, CNN, loss function

**Received:** December 8, 2020

*Face recognition has important value in real life. In this study, the application of the deep learning method in the field of face recognition was studied. The structure of LeNet-5 in convolutional neural network (CNN) was selected and improved; based on it, a face recognition method was designed. The performance of the method was analyzed taking CelebA as training set and LEW as testing set. The results showed that the improved LeNet-5 model which took A-softmax Loss as loss function not only had shorter training time, but also had higher recognition accuracy, its accuracy increased with the increase of sample size, and the highest accuracy rate reached 97.9%. The experimental results showed that the face recognition method designed in this study had good performance in large data background as it could effectively reduce the running time of the algorithm and improve the recognition accuracy. This study proves the reliability of deep learning methods such as CNN in face recognition, which is conducive to the further development of face recognition technology.*

*Povzetek: Opisano je prepoznavanje obrazov z metodami globokih nevronskih mrež in z velikimi podatki.*

## 1 Introduction

With the development of computer technology and in the context of big data, people pay more attention to issues such as data security and personal privacy, and the social requirements for human identification are also increasing. Traditional identification methods based on identity cards and passwords have low reliability because they are easy to be counterfeited and lost. Therefore, biometric identification technologies such as fingerprints and voices have been widely recognized [1]. Face recognition is a kind of biometric recognition, which has attracted more and more research and attention. However, due to the difference of face pose and illumination, face recognition is difficult [2]. The deep learning method has excellent performance in face recognition, especially in big data processing [3], and relevant research is also deepening. Ding et al. [4] studied the recognition of face images with severe noise. Based on the deep neural network, an anti-noise network was designed, and the reliability of the network in face recognition with noise was proved by experiments. Lu et al. [5] proposed a deeply coupled ResNet model, which was composed of a relay network and two branch networks. It could extract various possible resolutions of images, and the reality of the model was proved by experiments in LFW and SCface databases. Jiang et al. [6] designed an unsupervised deep learning network by combining 2-D Gabor filter with PCA to improve the computing speed through short binary hashing and then proved the excellent performance of this method by testing in face database. Singh et al. [7] applied convolutional neural network (CNN) to neonatal recognition and found that CNN had a good accuracy in neonatal recognition compared with conventional technology and CNN with two convolution layers and one hidden layer had the highest accuracy. In this study, deep learning was analyzed. Based on CNN in deep learning, a face recognition method was designed. The reliability of the method was proved by LFW data set, which provides some theoretical support for the further application of deep learning in face recognition.

## 2 Face recognition

Face recognition refers to extracting feature information from static or dynamic images collected by computer and then analyzing and matching to realize identity recognition. Compared with other biometric methods, face recognition image acquisition is more convenient, with rich personal characteristics, high recognition degree and good interaction. It has been widely used in surveillance video, intelligent consumption [8], criminal investigation [9] and so on.

Traditional face recognition methods include geometric features, template matching and so on, but there are also some shortcomings. Face feature extraction is a very important step in recognition, which has a great impact on the final results. In traditional recognition methods, feature extraction is mostly based on manual method. Under the background of massive data, the traditional recognition methods not only take a lot of time and energy, but also are difficult to recognize images because they are easily affected by illumination, occlusion and other factors. Deep learning can automatically extract

features, which is less affected by external factors, and it has been proved to have good recognition effect.

# 3 CNN

## 3.1 Overview of CNN algorithm

CNN is a common model of deep learning. Its basic structure is shown in Figure 1.

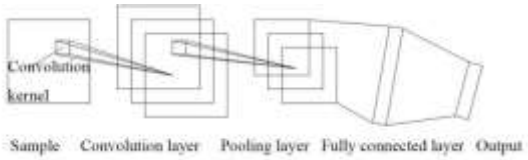



Figure 1: The structure of CNN.

(1) Convolutional layer
Convolution layer is the core component of CNN. It extracts image features by convolution operation, generates different feature maps by different convolution kernels and superimposes them to obtain various features of input image. Its output calculation method is:

$$y_j^l = f\left(\sum_{i=1}^{N_j^{l-1}} w_{i,j} \otimes x_i^{l-1} + b_j^l\right), j = 1,2,\cdots,m$$

where $l$ represents the current number of layer, $w$ represents the convolution kernel weight matrix, $x_i^{l-1}$ represents the output characteristic pattern matrix, $f$ represents an activation function, $\otimes$ represents convolution operations, and $b_j^l$ represents the offset of the $j$ -th characteristic pattern of the $l$ -th layer.

(2) Pooling layer
The role of the pooling layer is to compress data and reduce the amount of computation. There are two common methods, average pooling and maximum pooling. Figure 2 shows an example of maximum pooling. The size of image is 4×4, the size of pooling window is 2×2, and the step length of maximum pooling operation is 2. In the first pooling window, the values are 5, 7, 9 and 2 and the maximum value is 9; thus the maximum pooling result can be obtained

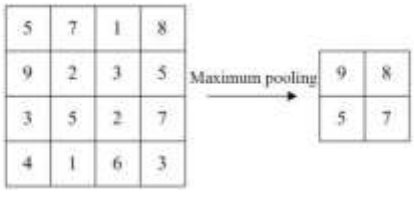



Figure 2: Maximum pooling.

by traversing the whole image.

(3) Fully connected layer
Fully connected layer plays the role of classification, and its calculation formula is:

$$\delta_j^l = f\left(\sum_{i=1}^{n} x_i^{l-1} w_{ij}^l + b_j^l\right)$$

where $l$ represents the current level, $n$ represents the number of neurons, $w$ represents weights, $b_j^l$ represents offset, and $f$ represents an activation function.

## 3.2 Training process of CNN

The training process of CNN can be divided into two stages:

(1) Forward propagation
A sample $(X, Y_p)$ is selected from the sample set, and $X$ is input into CNN.
Actual output $O_p$ of CNN is calculated.

(2) Reverse propagation
(1) The error between actual output $O_p$ and expected output $Y_p$ is calculated.
(2) The error is reversely propagated, weight matrix is adjusted, and parameters are optimized.

# 4 Face recognition based on deep learning

## 4.1 Experimental environment

The experiment was carried out on Ubuntu 16.04 operating system. The program was written in C++ language and Python language. The training and testing of CNN model was realized by Caffe framework, which supports GPU acceleration, runs faster and operates more simply.

## 4.2 Experimental data set

At present, data sets commonly used in face recognition include CAS-PEAL, CASIA-WebFace, LFW, MSCeleb, CelebA and so on. In this study, CelebA was selected as the experimental training set, and LEW was used as the testing set. CelebA can train the model well as it includes 200,000 face images of 10,177 people and there are changes in expression, posture, occlusion and illumination. LFW which has been widely used in the performance analysis of face recognition algorithms includes 13,233 images, a total of 6000 face combinations.

## 4.3 Data preprocessing

The main task of data preprocessing is face alignment. As the face image is partly inclined (Figure 3), the difficulty of recognition increases. Therefore, in order to obtain better recognition effect, image alignment is needed. The face images obtained after alignment are shown in Figure 4.

Figure 3: Face images.

Figure 4: Face images after preprocessing.

### 4.4 Improved LeNet-5

LeNet-5 is one of the most representative structures in CNN [10]. In order to improve the recognition performance of the network, the structure of LeNet-5 was improved in this study. Five convolution layers, four pooling layers and one fully connected layer were used. The specific parameters of each layer are shown in Table 1.

| Type | Convolution kernel | Number of characteristic patterns | Number of neurons |
|---|---|---|---|
| Convolution layer 1 | 5×5 | 16 | 16128 |
| Pooling layer 1 | 2×2 | 16 | 4032 |
| Convolution layer 2 | 4×4 | 32 | 1536 |
| Pooling layer 2 | 2×2 | 32 | 384 |
| Convolution layer 3 | 3×3 | 64 | 128 |
| Convolution layer 4 | 6×6 | 16 | 4032 |
| Pooling layer 3 | 2×2 | 16 | 1008 |
| Convolution layer 5 | 5×5 | 32 | 480 |
| Pooling layer 4 | 2×2 | 32 | 256 |
| Fully connected layer | - | - | 192 |

Table 1: Improved LeNet-5.

In order to improve the training speed of the algorithm, an improved ReLU function, LReLU, was used as the activation function of the model:

$$\mathrm{LReLU}(y) = \begin{cases} y, if(y > 0) \\ ay, if(y \le 0) \end{cases}$$

where $a$ represents a small constant, so that the function is not zero when the input is negative, preventing neuron necrosis.

There were two choices of loss function for the model: Softmax and A-softmax Loss:

(1) Softmax: For input $x$, it is divided into $k$ classes, then the probability of sample belonging to class $i$ can be expressed as:

$$g_\theta(x^{(x)}) = \begin{bmatrix} p(y^{(i)} = 1 | x^{(i)}; \theta) \\ p(y^{(i)} = 2 | x^{(i)}; \theta) \\ \cdots \\ p(y^{(i)} = k | x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \cdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix}$$

where $g_e(x)$ is a hypothetical functions and $\theta_i$ is a model parameter.

(2) A-softmax Loss: A-softmax Loss is an improvement of Softmax, which introduces angular distance and angular margin, and its expression is:

$$L_{ang} = \frac{1}{N} \sum_i - \log \left( \frac{e^{\|x_i\| \cos m\theta_{y_i}, i}}{e^{\|x_i\| \cos m\theta_{y_i}, i} + \sum_{j \neq y_i} e^{\|x_i\| \cos m\theta_j, i}} \right),$$

where $m$ represents an integer, which is used for controlling the angular distance.

## 5 Experimental results

Images of 100 people were selected from CelebA to train the model, ten images each people. The training time of different models is shown in Table 2.

| CNN model | Loss function | Training time |
|---|---|---|
| LeNet-5 | Softmax | 59.27 s |
| LeNet-5 | A-softmax Loss | 57.46 s |
| Improved LeNet-5 | Softmax | 45.39 s |
| Improved LeNet-5 | A-softmax Loss | 42.18 s |

Table 2: Comparison of training time of models.

It was found from Figure 2 that the training time of LeNet-5 model was longer than that of the improved LeNet-5 model when using the same samples. In the same CNN model, the training time of the model which used A-softmax Loss as the loss function was shorter than that of the model which used Softmax function, and the training time of the improved LeNet-5 model with A-softmax Loss as the loss function was the least.

Taking A-softmax Loss as the loss function, two CNN models were tested using LFW data sets. 100 pairs, 500 pairs, 1000 pairs and 2000 pairs of matched face images were taken as positive samples; as shown in Figure 5, the two images matched each other, which was called a pair of positive samples. Mismatched face images were taken as negative samples; as shown in Figure 6, the two images did not match, which was called a pair of negative samples.The recognition results of the model can be divided into four cases, as shown in Table 3.

The recognition accuracy of the model = (TP+TN)/ the total number of samples.

Under different number of samples, the recognition accuracy of the two models is shown in Figure 7.

It was found from Figure 7 that the recognition accuracy of the model increased with the increase f the

Figure 5: An example of positive sample.

Figure 6: An example of negative sample.

| | Identified as positive samples | Identified as negative samples |
|---|---|---|
| Actual positive sample | True Positive (TP) | False Positive (FP) |
| Actual negative sample | False Negative (FN) | True Negative (TN) |

Table 3: Classification of recognition results.

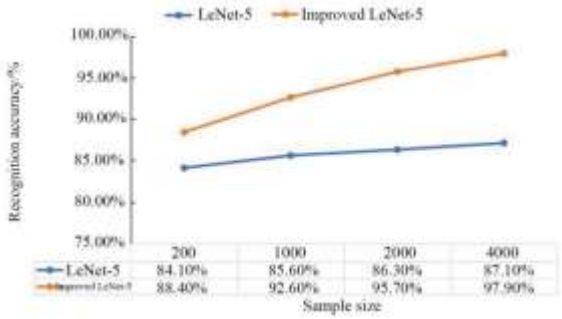



Figure 7: Comparison of recognition accuracy between models.

sample size, which showed that CNN model had excellent performance in recognizing massive face data and could accurately recognize large-scale data. From the comparison of the two models, it was found that the accuracy of the improved LeNet-5 was higher than that of LeNet-5. When the sample size was 4000, the recognition accuracy of LeNet-5 was 87.1%, while that of the improved LeNet-5 was 97.9%. The results showed that the improved CNN model could extract face features more comprehensively and obtain better recognition effect.

# 6 Discussion and conclusion

Deep learning is an important part of machine learning. It is based on big data and can automatically extract feature information from massive data by certain algorithms instead of traditional manual feature acquisition. It has higher accuracy than shallow learning and better performance in dealing with non-linear problems. It has shown great advantages in fields such as computer vision and semantic analysis. CNN is one of the deep learning methods, which has been widely used in object recognition and detection. With the support of massive data, face recognition based on CNN has excellent performance [11].

In this study, CNN was analyzed firstly. Traditional recognition methods, such SVM [12], can only extract shallow features when extracting image features, which is easily affected by other factors, and the recognition rate is not high. Deep learning methods such as CNN can extract abstract and conceptual features in depth [13], which is less disturbed by illumination, gesture and expression. CNN can extract multiple image features by convolution operation, then reduce the dimension by pooling layer to reduce the amount of calculation, and finally classify them. Based on LeNet-5 in CNN, the network structure was improved to make it more suitable for face image processing. Then, the improved ReLU function, LReLU, was used as activation function, and the influence of loss function on the performance of the model was analyzed. In the experiment, CelebA was used as training set to train the model, and then LEW was used as testing set to test the performance. The results showed that the improved LeNet-5 model using A-softmax Loss had shorter training time among LeNet-5 models using softmax and A-software Loss as the loss function and the improved LeNet-5 models, which showed that it had higher convergence speed. Then in the processing of LFW testing set, A-softmax Loss was used as the loss function, and the recognition accuracy of the improved LeNet-5 was significantly higher than that of LeNet-5. The recognition rate of the two models increased with the increase of sample size, and the gap between the two models increased as well. When the sample size was 4000, the recognition accuracy of LeNet-5 was 87.1%, while that of the improved LeNet-5 was 97.9%.

In summary, the face recognition method designed in this study has short training time and high recognition accuracy. It has excellent performance when facing a large number of face images. The reliability of deep learning methods such as CNN is proved, which makes some contributions to their further application.

# 7 References

[1] Galbally J, Marcel S, Fierrez J (2014). Biometric Antispoofing Methods: A Survey in Face Recognition. IEEE Access, 2, pp. 1530-1552. https://doi.org/10.1109/ACCESS.2014.2381273.

[2] Zhang K, Zhang Z, Li Z, et al (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 23, pp. 1499-1503. https://doi.org/10.1109/LSP.2016.2603342.

[3] Pang SC, Yu Z (2015). Face recognition: a novel deep learning approach. Journal of Optical Technology C/c of Opticheskii Zhurnal, 82, pp. 237.

[4] Ding Y, Cheng Y, Cheng X, et al (2017). Noise-resistant network: a deep-learning method for face

recognition under noise. Eurasip Journal on Image & Video Processing, 2017, pp. 43.
[5] Lu Z, Jiang X, Kot C (2018). Deep Coupled ResNet for Low-Resolution Face Recognition. IEEE Signal Processing Letters, pp. 1-1. https://doi.org/10.1109/LSP.2018.2810121.
[6] Jiang M, Lu R, Kong J, et al (2017). GB (2D) 2 PCA-based convolutional network for face recognition. Neuroreha, 06, pp. 131-135.
[7] Singh R, Om H (2017). Newborn face recognition using deep convolutional neural network. Multimedia Tools & Applications, 76, pp. 1-11. https://doi.org/10.1007/s11042-016-4342-x.
[8] Smith DF, Wiliem A, Lovell BC (2015). Face Recognition on Consumer Devices: Reflections on Replay Attacks. IEEE Transactions on Information Forensics and Security, 10, pp. 736-745. https://doi.org/10.1109/TIFS.2015.2398819.
[9] Ghiass RS, Arandjelovic O, Bendada H, et al (2014). Infrared face recognition: a comprehensive review of methodologies and databases. Pattern Recognition, 47, pp. 2807-2824. https://doi.org/10.1016/j.patcog.2014.03.015.
[10] Zhao ZH, Yang SP, Ma ZQ (2010). License Plate Character Recognition Based on Convolutional Neural Network LeNet-5. Journal of System Simulation, 22, pp. 638-641. https://doi.org/10.3724/SP.J.1187.2010.00953.
[11] Wu W, Yin Y, Wang X, et al (2018). Face Detection With Different Scales Based on Faster R-CNN. IEEE Transactions on Cybernetics, PP, pp. 1-12. https://doi.org/10.1109/TCYB.2018.2859482.
[12] Zhang L, Zhou WD, Li FZ (2015). Kernel sparse representation-based classifier ensemble for face recognition. Multimedia Tools & Applications, 74, pp. 123-137. https://doi.org/10.1007/s11042-013-1457-1.
[13] Li Y, Lu Z, Jing L, et al (2018). Improving Deep Learning Feature with Facial Texture Feature for Face Recognition. Wireless Personal Communications, pp. 1-12.

https://doi.org/10.31449/inf.v44i4.3253 Informatica **44** (2020) 497–505 **497**

# Data Protection Impact Assessment Case Study for a Research Project Using Artificial Intelligence on Patient Data

Gizem Gültekin Várkonyi
International and Regional Studies Institute, Faculty of Law, University of Szeged
6720 Szeged, Tisza Lajos krt. 54, Hungary
E-mail: gizemgv@juris.u-szeged.hu

Anton Gradišek
Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
E-mail: anton.gradisek@ijs.si

**Keywords:** data protection, DPIA, GDPR, Artificial Intelligence, medical data

**Received:** July 28, 2020

*Advances in artificial intelligence, smart sensors, data mining, and other fields of ICT have resulted in a plethora of research projects aimed at harnessing these technologies, for example to generate new knowledge about diseases, to develop systems for better management of chronic diseases, and to assist the elderly with independent living. While the algorithms themselves can be developed using anonymized or synthetic data, conducting a pilot study is often one of the key components of a research project, and such studies unavoidably involve actual users with their personal data. Although one of the derogations stipulated in Article 89 of the GDPR is related to the data processed for scientific purposes, the GDPR still is applicable to that processing in a broader interpretation. The computer scientists and engineers working in research projects may not always be fully familiar with all the details of the GDPR, a close collaboration with a lawyer specialized in the European data protection legislation is highly beneficial for the success of a project. In this paper, we consider a hypothetical research project developed by an engineer dealing with sensitive personal data and a lawyer conducting Data Protection Impact Assessment to ensure legality and quality of the research project.*

*Povzetek: Prispevek obravnava varstvo osebnih podatkov pri uporabi metod umetne inteligence za analizo pacientov.*

## 1 Introduction

In general, there are two ways to look at artificial intelligence (AI) dealing with personal data. On one hand, it offers great benefits for the users, if used correctly. For example, AI-enabled health care technologies could predict the treatment of diseases 75% better, and could reduce the clinical errors 2/3 at the clinics using AI compared to the clinics that do not [1]. On the other hand, the improper handling of personal data can quickly lead to abuse, sharing sensitive information, or other problems (unwanted disclosure, complex legal procedures, high amount of fines, etc.), therefore it has to be handled with the utmost care. In this paper, we will focus on the medical applications, such as the analysis of sensor data to help patients with chronic diseases manage their condition and improve the quality of life, or to help the elderly with independent living by providing safety features and improved communication channels.

Developing a product for the target population, for example people with diabetes, chronic heart failure, obesity, dementia, skin cancer, etc., typically starts with a research project, either in a company or within a consortium of research institutions and hospitals. One of the key components of such a project is collecting substantial amounts of data in a pilot study, with participants that resemble the target audience for the final product. When planning the pilot study, researchers enter a slippery terrain of dealing with personal data, as the participants are providing their own data for the purpose of the study. For the purpose of this paper, we will have a closer look at the medical data encapsulating three forms; general medical data provided by the medical doctor responsible for the participant, lifestyle data collected by either wearable or stationary sensors, and self-reported data that is obtained via questionnaires that the participants fill. At this stage, we only look at the data from the point of view of the hypothetical research project, and do not take into account the implications that are brought by potential commercial exploitation of the findings.

Right to data protection is one of the fundamental rights recognized in most of the European legislation, mainly in the Charter of Fundamental Rights and the General Data Protection Regulation (GDPR). The GDPR entered into force on the 25th of May 2018 replacing the Directive 95/46/EC based on two main aims: ensuring uniform data protection rights and rules EU-wide and towards data controllers, and keeping up with the technological developments challenging efficient

protection of personal data [2]. The effect of technology pointed out the need for more proactive ways to safeguard right to data protection and the GDPR mirrored this need by introducing a risk-based approach entrusted in the Article 35 of the GDPR introducing the Data Protection Impact Assessment (DPIA). As such, DPIA ensures data controllers comply with the GDPR requirements especially at an early stage of a new project. Those requirements could be specific to the right to data protection introduced in the GDPR such as the Article 25-Data Protection by Design, or to general principles that have already existed in European data protection legislation such as the principle of accountability. In fact, DPIAs are one of those ways for materializing and ensuring the accountability principle which has always been a legal compliance element and is now being utmostly challenged by the risks deriving from the new technologies [3].

The year 2018 was quite a productive year for the European Commission (EC) in terms of regulation of AI in the EU. Firstly, the EC published EU's AI Strategy [4] and then the EU's Coordinated Plan on AI [5] which both focused on the importance of system design which should be human-centric and trust-gaining. Both documents point out the data protection and privacy concerns as a problem, and suggest that legal compliance together with ethical system design is at the utmost importance to gain trust of AI users which then could boost the AI developments in the EU. The DPIA requirement embedded in the GDPR is such a tool that could be used as a proof before the users to gain their trust towards the AI system. For this reason, we think that the DPIA is an essential for any AI project planned to be targeted in the EU should be considered, if the project stakeholders aim at fulfilling both legal compliance and gaining individuals' trust. Individuals, who then might be data subjects in case they contribute to the AI system development in the training phase with their data, will enjoy the possibility of exercising their rights explicitly presented them by the DPIA output. Especially, they could receive descriptions specific to a systematic automated decision making processes since DPIA also aims to identify the logic involved in the algorithm as well as the significance of the consequences of the algorithmic evaluations [6]. However, yet there is no standard set for conducting a DPIA by the law-maker, as well as there is a lack of experience in practice since the GDPR is quite a young legislation. Specific to the AI driven research project collecting and processing personal medical data, there is no example existed in the literature illustrating a DPIA implementation, even though there are DPIA applications specific to AI implementations such as one for assessing the risks deriving from AI use in decision-making [7], or the work offering a roadmap for assessing the social and ethical impact of AI [8]. Furthermore, lack of specific examples to the DPIA on a certain technology, such as smart cities, may cause wrong identification of the risks which then may hinder data controller's full legal compliance [9]. In this paper, we aim to fill this gap with an example DPIA implementation on a hypothetical research project aiming to develop an AI system. Following content of the paper will be focusing on illustrating the legal foundations of the DPIA as in the GDPR in Section 2. Next, the hypothetical research case will be introduced which will then be followed by the DPIA practice in Section 3. According to the analysis conducted in the Section 3, the paper identifies three assessment titles specific to the AI projects: data specific assessment, data subject specific assessment, and project specific assessment. Section 4 presents a conclusion and a set of recommendations deriving from the outputs of the analysis conducted.

## 2 Data protection impact assessment in the GDPR

Article 35 of the GDPR does not provide an explicit description for the DPIA, however, Article 29 WP's guideline on the DPIA provides the following definition: "A DPIA is a process designed to describe the processing, assess its necessity and proportionality and help manage the risks to the rights and freedoms of natural persons resulting from the processing of personal data."

According to this definition, and in a narrower sense, the DPIA is a process consisting of several other sub-processes to describe the risks and assess the legality of the system in terms of data protection. These risks could be related to system security, system design, implementation, administration and development on a further run. The aim of the DPIA is to take appropriate safeguards to minimize the risks, if impossible to eliminate all. DPIA is not a simple one-time reporting activity, it is an ongoing process that should be continuously carried out during the lifetime of a project, therefore DPIA should always be monitored and updated [10].

It is the data controllers' responsibility to convey a DPIA, but in fact, the GDPR does not assign them an obligation to carry out a DPIA for every data processing activity. DPIA should be carried out when the data processing activity is likely to constitute a "high risk" to the rights and freedoms of natural persons (e.g. users of an AI service who both benefit from the service and contribute to it with their data), as the Article 35 (1) refers. The existence of automated decision-making tools applicable on personal data, and processing sensitive data such as medical data are some of the criteria conceptualizing the term high risk according to the Article 35 (3) of the GDPR. In addition, there are several guidelines published so far by the National Supervisory Authorities aiming to create a list of data processing activities likely to result in a high risk. Currently, all the National Supervisory Authorities of the 27 Member States have formed such a list. The European Data Protection Board assessed and delivered its opinions on each list to ensure consistent implementation of the rules in the EU. These lists could be the first sources for the data controllers to decide about the necessity of the DPIA for a certain project [11].

After determining the necessity to conduct a DPIA, the next step should be assessing the severity and likelihood of the risks which would come forward based on the data controller's own assessment. Although there is

no standard specified for how to convey a DPIA, failure to conduct a right assessment raises a risk for the data controllers; they may face several sanctions, especially financial penalties. Apart from that, conducting a right DPIA would be beneficial for the data controllers not only from the legal and the financial point of view. Wright [12] lists these benefits and refers that a DPIA could help data controllers to avoid implementing irrelevant solutions from the beginning of the project which may refer to assessing the technical feasibility of the system in parallel with the legal compliance. Therefore, the DPIA could help data controllers to save time and money. It also prevents the companies from losing their reputation (or from the scandals, as occurred with Cambridge Analytica, Equifax, Facebook, etc.). Finally, a DPIA document can be a trustworthy source which could be used as evidence before the public, and the related authorities to prove the data controller's respect to privacy.

When planning a research project, the DPIA shall not be conducted neither after launching nor during the implementation in order to ensure proactive measures. Specific to our project in the present article, the DPIA should be conducted based on two legal obligations as provided by the GDPR. Firstly, the Article 35 (3) point (a) of the GDPR clearly indicates that those data controllers that are using automated tools to evaluate personal aspects of natural persons, including profiling, are required to conduct a DPIA. Secondly, as the (b) point of the said article indicates, processing special categories of data also requires data controllers to conduct a DPIA. Medical data that will be evaluated in the project, as indicated before, is classified under the special data categories. Furthermore, the project focuses on developing an AI-based system that includes an automated decision making system (the AI software itself together with the algorithms to be developed) with profiling tools (surveys and hardware equipment). Based on these statements, it is clear that a DPIA must be conveyed by the data controller of the hypothetical project to see the risks and safeguard these risks in line with necessary tools. These tools might be either organizational or technical tools that could help mitigating the risks. The safeguards will be presented in the analysis part of the case study below.

# 3 Case study

In this paper, we present a step-by-step DPIA practice for our research project aiming to develop an AI-based healthcare software. While such approach is not a new in the literature and there are pieces of publications assessing the data protection risks of the real projects similar to ours [13], these are few in numbers and the DPIA really lacks in practice since data controllers usually do not prefer to publish their DPIA reports, as it is not required by law. We developed the idea of presenting a hypothetical research project that could exhibit a part of a realistic work and could contribute to the DPIA practices in the literature. Moreover, there are few examples specifically evaluating the data protection impact of an AI based software project. Following, we present the details of this AI software project and exhibit the simple DPIA elements that we created from several resources available on how to conduct a DPIA such as the Information Commissioner's Office guidelines[3], Deutsche Telekom's practices [14], French Data Protection Authority (CNIL) guidelines [15], and Article 29 Working Party guidelines [16].

## 3.1 Summary of the hypothetical research project

The goal of our research project is to discover new knowledge about a particular chronic disease and to develop a coaching system that will allow the patients for better management of their condition. In order to obtain sufficient amounts of data to build a personalized coaching tool, the researchers need to obtain various types of data from, say, 200 patients with the chronic condition, through a pilot study. In the pilot, the users are equipped with a smart wearable device that records the amount and intensity of daily activities, on aggregate. Such devices may come in the form of wristbands, smart watches, smartphones placed on various spots on the body, chest harness to monitor a simple electrocardiogram and breathing rate, or dedicated pendants. The interaction with the user likely takes place through a tablet or a smartphone. In addition, the application occasionally asks the patient questions related to their psychological state, as well as about some of their habits, such as smoking, consumption of alcohol, and about the dietary preferences. Medical doctors who recruited the patients for the pilots provide relevant data about the medical history of the patients, such as the timeline of their condition, the severity, and the medication or medical devices that the patient is using. The data is then processed using computer algorithms which find novel relations between various parameters, such as the effects of lifestyle on the expression of the condition, or how particular treatments help different patients best. Deeper information on the project will be gained during the DPIA, since it would be quite risky to first finalize the project details, and to conduct a DPIA afterwards [15].

Types of data subjected to the processing activity in the research project are qualified as special categories of data, also known as sensitive data, according to the GDPR. Sensitive data needs stricter protection, for example, the data subject's explicit consent should be obtained before launching the project. In order to obtain a valid explicit consent, the data controller must be able to present precise and specific information on the life-cycle of the data to be processed during the project. In addition, processing sensitive data may fall under the high-risk data processing category according to the provisions of the GDPR, therefore the data controller should conduct a DPIA prior to launching the project.

## 3.2 Requirements for a DPIA

The algorithm planned within the AI based healthcare software project is going to enable collecting data subjects' sensitive data based on profiling and processing that data. In addition, a large amount of data will be collected for feeding the algorithm conveying a risk for

data subjects, basically, the data subjects may cause loss of significant control of their own data. Based on these inputs, the project may reveal risks for rights and freedoms of the data subjects involved, if these risks are not mitigated. Therefore, it is a clear obligation for the data controller to conduct a DPIA and identify the risk categories with the planned mitigations when necessary. The following part shall present an assessment part of the actual DPIA since we skip the preparation phase of a regular process that includes planning, document collection, consultations with the stakeholders, etc. [10], which is not of particular interest for this paper.

## 3.3 DPIA for an AI-based healthcare research project

In this section, we conduct a DPIA on our research project. Several components are likely to be encountered while assessing any healthcare-related project, though each project has its own peculiarities. Therefore, this assessment is not universal, but can be viewed as an example for the engineers who are not deeply involved with the GDPR and who are looking for guidance to start with the preparation of the document.

We recommend that any DPIA should be conveyed under the supervision of a GDPR expert or a lawyer. As indicated before, the structure of the following section rely on several papers generated by the authorities guiding data controllers on how to conduct a DPIA. The questions and the answers referred to in this section stem from the author's own experiences, therefore the following DPIA is giving a lawyer's and an engineer's point of view. The structure of the below DPIA example is as follows: data specific assessment (DSA), data subject specific assessment (DSSA), and project specific assessment (PSA).

### 3.3.1 Data specific assessment

The DSA is chosen to be processed on the first hand because such an approach would shape the outcomes of the two other assessment groups. The DSA is the procedure where the data to be used in the AI project should be introduced very specifically in order to comply with the basic rules of the GDPR, mainly, the purpose limitation, transparency, accuracy, data minimization, and consent. It should be kept in mind that one of the requirements to be ensuring a valid consent is identifying the concrete data list together with the planned process of that data in the frame of a research project. Based on these statements, we propose the following questions placed in the Table 1. to be considered as part of the DSA.

The DSA questions raised here are related to the life-cycle of the data in the research project. Three types of data are planned to be collected during the research and all the types are clearly defined. The boundaries of the medical, activity, and self-reported data are reported in line with the data minimization and purpose limitation principles. Sources of the data are also clear and limited. It is crucial to note the responsible person for collection and processing of the data and the retention period is calculated. The data retention period should be followed without a prejudice to the Article 17 of the GDPR ensuring data subjects' right to erasure. This project does not aim at reusing data at the moment giving as a reason that there are several problems standing before data reusing rules and personal data protection legislation [17] refraining the project team from opening the research data for other purposes. However, there is a challenge identified with the models which will be reused, since it is well-known that with an intended attack, the data in training sets could be revealed [18], [19], [20]. This risk is mitigated with security measures and methods ensuring privacy specific within the AI models. In addition, AI models are not regulated under the GDPR except with the general rules such as Privacy by Design. More guidelines about protection of data in AI models could be delivered either from the European or from national data protection authorities. Finally, the 6th question in the table deriving from Article 13 and Article 15 of the GDPR raises concerns for the project team on how to ensure the full compliance with the GDPR if it is not possible to foresee the algorithm to be used from the beginning of the data processing. This problem is based on the technical construction of the AI and the legal uncertainty on the meaning of the logic-involved within the GDPR. If the term logic-involved means the planned algorithm to be developed, then it is not possible to give a concrete answer from the beginning of the project. If the rights vested in the Article 13 and Article 15 of the GDPR are the reactive rights rather than a proactive, then the project team can explain the logic of the algorithm, later. In any case, this risk is mitigated with a clear indication in the consent paper informing the data subjects about the concern.

### 3.3.2 Data Subject Specific Assessment

The DSSA should explain all the details of how the data controller ensures the rights of the data subjects and protects their informational self-determination right. The key point in this assessment is to gain trust of data subjects as required by law and ethics. The DSSA questions are mostly about how the data subjects rights will be ensured during the project. It is highly recommended to work with a Data Protection Officer in line with the Article 37-1 (b) of the GDPR, since data processing activities in this project require regular and systematic monitoring of data subjects on a large scale". The DPO whose duty is to consistently follow and ensure the communication between the data subjects and the project team, among the other tasks drawn in the Article 39 of the GDPR, could be chosen among the project team members, or to be contracted in line with the qualifications indicated in the GDPR. In this project, the DPO is the responsible person to guide the project team about the data subjects' requests and their fulfilment and is a lawyer specialized in data protection law. For this reason, we recommend the project team to work with a lawyer or a data protection expert from the beginning of the project development.

The importance of the 3rd question is vested in the clarity of the consent statement that shall be read and understood by each data subject. The project team could plan to involve the data subjects' opinion on the draft

| | |
|---|---|
| 1. What type of data, in what format, and in what scale will be processed? | Medical data (age, gender, medical history of other comorbidities, medications, clinical data related to the condition)<br>Activity data (aggregated amount of physical activity per day and the physical activity, which is already defined in the course of the research and in the consent statement)<br>Self-reported data (questionnaires prepared by the research team in collaboration with physicians)<br>The study is limited to about 200 patients and the number will not exceed this. |
| 2. What are the sources of the data? What measures are taken to ensure security of the sources of the data? | Medical data come from the treating physicians. Activity and self-reported data come from the patients through a smartphone application. All data transfers are secured with encryption algorithms and immediately anonymized. |
| 3. Who will collect the data? Who will have an access to the data? | The medical data will be collected by their treating physicians. The activity and self-reported data will be collected through an application individually from each user.<br>The access to the data will be structured hierarchically, with different partners having access to different types of the data, but only the treating physicians will know the identities of the patients, as this is unavoidable. All other partners will only access anonymized data. |
| 4. Will the data be reused for another purpose in future? | The data will not be reused. What may be reused are the models that will be obtained by training the algorithms on the data. The models are protected with security and differential privacy measures against data revelation. |
| 5. How long will the data be processed? Where and until it will be stored? | The data will be processed during the duration of the project, which is 3 years. It will be then stored for another 5 years for potential purposes related to the research within the scope of the project. It will be stored on a secure offline server physically located at one of the partner organizations. After that, all personal data will be deleted. |
| 6. What technology will be used to process the data? | Before the data becomes available, it is difficult to answer this question. Typically, the algorithms used in such studies include decision trees, support vector machines, or different types of neural networks. |
| 7. Who is responsible for the security of the data (in storage and during the collection)? | There is a dedicated engineer in the project team who is responsible for data security and storage. |

Table 1: DSA questions for a DPIA and the corresponding answers regarding the project.

consent statement, and shape it in accordance with their feedback. Consultation with data subjects will also help the project team to know about their concerns, and mitigating their concerns would contribute the project to be more GDPR-friendly. The project team could also plan to make a half-day informative meeting with the data subjects on the project's essences and we present them the current DPIA. Data subjects are ensured with tools that could help them to withdraw their consent without an obstacle. Since the research is focusing on creation of the model, the users can ask for data deletion at any point while the development phase is ongoing. Once the models reach the final version, they will not contain any personal data. They are also given tools to download their data to be collected during the project and can exercise their right to data portability. It is planned that, since the development phase will take 12 months, the system will send automatic consent reminders every 3 months. Besides all these safeguards, data minimization is guaranteed with specific privacy setting interfaces embedded in the wearables or the sensors which will be used for lifestyle data collection. There will be a separate training designated for how to use the device and the privacy settings. The 4th question points out the well-known black-box debates that is weakening the intervention capability of the project team on the decision given by an algorithm, if the data subject wishes to exercise his or her right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision, as the Article 22 of the GDPR stresses. However, as this is a pilot project, the decisions are neither a final service nor a product, therefore they are only recommendations. Additionally, the project team guarantees the data subjects to express their own point of views (i.e. feedback) which can help the team to better develop or to modify the system mistakes.

| | |
|---|---|
| 1. What is the legal basis for processing data? | Article 6 (a) of the GDPR (data processing based on data subject's consent). |
| 2. What is the purpose of the project? | The purpose of the project is to generate new knowledge about a chronic disease and to create a coaching platform that will be assisting the patients with management of their condition, thus greatly improving their quality of life. |
| 3. What is the purpose of data processing? Are the purposes indicated in (2) fully in line with each other? | The purpose of the data processing is twofold:<br>• to seek for new relations in the data which will lead to better understanding of the condition<br>• to train personalized models that will make the experience for individual patients better |
| 4. What is the expected benefit of the project for the data subjects, for the data controller, and for the society? | Overall benefit of the project will be for the individuals and for the society helping to develop a system that will be beneficial for them in future, for improving the quality of life, and reducing the burden on the health system for the society |
| 5. How many stakeholders are involved with this project? Who is the data controller and how to identify the data controller? | If there are more than one stakeholders who can process the data, it is recommended to designate one of them as a contact point. The contact point's availability information should be easily accessible by the patient (address, phone, e-mail, preferable channel for communication, and the information of the DPO). |
| 6. Are there any data processors? If yes, does the data controller have evidence of data processors GDPR compliance? | All personnel dealing with the data (data processors) are required to sign the GDPR compliance document, provided by the project leading institution. |
| 7. How does the data controller ensure security of the data during collecting, processing, storing, and removing the data? | Collection: anonymization at the source, encrypted protocols for data transfer<br>Processing: dealing with anonymized data only, no external access to the database<br>Storing: offline storage at secure location after the project has ended<br>Removing: authorized person deletes the data 5 years after the end of the project, if previously not requested by the patient<br>Hardware safety measures, including wearables security, is ensured by the in-built safety measures of the devices provided by the manufacturer that proves GDPR compliance. The devices will communicate with the data-collection platform using only the patient's ID. |

Table 3: PSA questions for a DPIA and the corresponding answers regarding the project

### 3.3.3 Project Specific Assessment

The PSA is the last part of our DPIA, presenting and explaining the legal basis for data processing, the project partners including the data controller, and the security measures that will be implemented to safeguard the data processed during the project. The security measures encompass a large and an important part of this assessment, therefore we present the security measures in a separate table.

The PSA table includes questions related to identification of the project purposes and legal basis as well as of the data controllers and processors. Data processing in this project is based on the data subjects consent and explicit consent where necessary. Data subjects are expected to participate in the project voluntarily, and they could make a decision about that participation based on the purposes of the project, expected outcomes and privacy statements to be provided for them. It is crucial for the project team to provide these information together with the explicit information on the identity of the data controller and other stakeholders, if there are any. In the PSA table, the second question raised some challenges while answering. The project aims at creating an AI-based healthcare software, however, since the data to be collected is Big Data, the project team is aware of the security risks and take necessary steps to ensure system security (see the Table. 4). Additionally, the project team ensures that, by design, the recommendation system will not work outside of the domain it was built for; therefore unpredictable outcomes are highly unlikely. In order to complete the administrative safeguards in this project, the project needs to present the necessary technical safeguards that are under the security measures.

| **Security risk** | **Security measure** |
|---|---|
| Data partitioning (in relation to the rest of the information system) | Hierarchical access to the data, user roles |
| Logical access control | Hierarchical access to the data, user roles |
| Traceability (logging) | Use of dedicated file monitoring software |
| Integrity monitoring | Use of dedicated file monitoring software |
| Archiving | Periodic archiving |
| Paper document security | Paper documents kept in a locked filing cabinet or similar |
| General security controls regarding the system in which the processing is carried out | Dedicated preventive control measures |
| Operating security | Dedicated preventive control measures |
| Clamping down on malicious software | Up-to-date malicious software removing tools installed |
| Managing workstations | Dedicated personnel |
| Website security | Standard measures for website security |
| Backups | Periodic backups, automated |
| Maintenance | Dedicated personnel |
| Security of computer channels (networks) | Encrypted communication, when dealing with external sources (receiving the data during the pilots) |
| Monitoring | Data protection officer |
| Physical access control | Institute's physical access policy |

Table 4: Security risks and measures (Extracted from [14], pp. 12-17).

The measures presented in the table are optional and may change depending on the project.

The final but an ongoing phase of the DPIA is the monitoring phase. Whenever there is a new element embedded in the project, and this element seems to change the balance of risk earlier assessed, the DPIA should be reviewed. This element could be involving a new data type in the algorithm or planning a commercial use of the algorithm. Bearing in mind the fact that ML and algorithms are referred to as entirely new technologies [3] and the growing amount of data together with a variety of hardware would raise the privacy risks [21], we suggest the project team to review the DPIA periodically.

# 4 Conclusion

Data Protection Impact Assessment is an integral part of any research project focusing on development of an AI algorithm with personal data. It should be conducted in the planning stage of the project and occasionally reviewed once the project is ongoing. This way, the project team, otherwise called data controllers, are able to identify the potential risks and find mitigation strategies for certain weak points. Last but not least, by conducting the DPIA, the project team fulfils the legal requirements, ensures higher trust of people involved, and avoids unforeseeable problems that might later occur.

It should be noted that there are automated general tools exist for the purpose of conducting DPIA [22], [23]. For instance, the open source DPIA tool freely offered by CNIL gives the possibility for the data controllers to compute the DPIA procedures with a step-by-step approach, and lets them make the risk calculation based on the weights identified for each risk labels, even though the risks are identified by the data controller manually. Users of the tool are guided with a set of questions categorized automatically and they could personalize the categories based on their needs. In the end, the tool gives the basics elements of the DPIA giving the data controllers opportunity to record also the mitigation records, but it would be highly beneficial for our project team to collaborate with a lawyer specialized in the data protection law, namely the GDPR, assisting them for using the tool. As demonstrated in this case study of a healthcare project

dealing with three types of personal data (medical, activity, and self-reported), DPIA is conducted through a series of steps, each of which addresses a different aspect of the data. We presented the DPIA in four tables, namely, Data Specific Assessment, Data Subject Specific Assessment, Project Specific Assessment, and Security risks and measures. Each table focuses on the particular aspect of the project and as close as it is compliant with the legal requirements. Step-by-step approach helped us to divide the data processing procedures and then evaluate each in detail. At the end of this assessment, it is safe to state that there is a low level of risk in this project, from the data protection point of view. This study aims to be an example for the community who is to plan an AI project and is looking for practical prior guidance to conduct a DPIA.

In this case study, we only focused on the use of personal data for the purpose of the research project. Clearly, successful research projects often continue with follow-up studies and eventually lead to commercial systems, in our case for example a smartphone-based coaching application for better management of a chronic disease. Commercial exploitation of research resulting from analysis of personal data opens a new series of questions. Can we commercially exploit the outcomes of this research project? Can a potential coaching application developed during the project be licensed to a commercial partner that will offer a subscription-based service? How to cover this in the agreement form that the participants sign before the beginning of the pilots? Such questions will be addressed in a future study.

## Acknowledgement

AG acknowledges the funding from the ERA PerMed project BATMAN, contract number C3330-20-252001. On Slovenian side, the project is funded by the Ministry of Education, Science, and Sport (MIZŠ).

# References

[1] "The AI effect: How artificial intelligence is making health care more human", [Online], study conducted by MIT Technology Review Insights and GE Healthcare, 2019. Accessed from: https://www.technologyreview.com/hub/ai-effect/ Last accessed: 20 April 2020.

[2] EDPS (2012). "Opinion of the European Data Protection Supervisor on the data protection reform package", (7 March 2012).

[3] ICO (2018). Accountability and governance: Data Protection Impact Assessments (DPIAs).

[4] European Commission (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe. COM (2018) 237 final.

[5] European Commission (2018) Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Coordinated Plan on Artificial Intelligence. COM (2018) 795 final.

[6] Kaminski, M. E. and Malgieri, G. (2019). Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. International Data Privacy Law, 2020, forthcoming, U of Colorado Law Legal Studies Research Paper No. 19-28, Available at SSRN: https://ssrn.com/abstract=3456224 or http://dx.doi.org/10.2139/ssrn.3456224.

[7] Ivanova Y. (2020) The Data Protection Impact Assessment as a Tool to Enforce Non-discriminatory AI. In: Antunes L., Naldi M., Italiano G., Rannenberg K., Drogkaris P. (eds) Privacy Technologies and Policy. APF 2020. Lecture Notes in Computer Science, vol 12121. Springer, Cham. https://doi.org/10.1007/978-3-030-55196-4_1.

[8] ECP Platform for the Information Society Netherlands (2019). Artificial Intelligence Impact Assessment. Last accessed: 21 April 2020. https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf.

[9] Vandercruysse, L., Buts, C., Dooms, M. (2020). A typology of Smart City services: The case of Data Protection Impact Assessment. Cities, 104, 102731. https://doi.org/https://doi.org/10.1016/j.cities.2020.102731.

[10] Wright, D. (2012). The state of the art in privacy impact assessment. Computer Law & Security Review, 28(1), 54–61. https://doi.org/https://doi.org/10.1016/j.clsr.2011.11.007.

[11] EDPS, Data Protection Impact Assessment: Accessed from: https://edpb.europa.eu/our-work-tools/our-documents/topic/data-protection-impact-assessment-dpia_en Last accessed: 20 April 2020.

[12] Wright, D. (2011). Should Privacy Impact Assessments Be Mandatory? Commun. ACM, 54(8), 121–131. https://doi.org/10.1145/1978542.1978568

[13] Horák, M., Stupka, V., & Husák, M. (2019). GDPR Compliance in Cybersecurity Software: A Case Study of DPIA in Information Sharing Platform. In Proceedings of the 14th International Conference on Availability, Reliability and Security. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3339252.3340516.

[14] Deutsche Telekom (2018). Guideline For the design of ai-supported business models, services and products at Deutsche Telekom in compliance with data privacy regulations.

[15] CNIL (2018). Privacy Impact Assessment Templates Accessed from: https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-2-en-templates.pdf Last accessed: 20 April 2020.

[16] Article 29 Working Party (2017). Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679.

[17] European Commission (2020). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A European strategy for data. COM (2020) 66 final.
[18] Aïvodji, U., Gambs, S., Ther, T. (2019). GAMIN: An Adversarial Approach to Black-Box Model Inversion. ArXiv, abs/1909.11835.
[19] Melis, L., Song, C., Cristofaro, E.D., Shmatikov, V. (2018). Exploiting Unintended Feature Leakage in Collaborative Learning. 2019 IEEE Symposium on Security and Privacy (SP), 691-706.
[20] Giuseppe A. et al. (2013). Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. arXiv:1306.4447v1.
[21] Chandra, S., Ray, S., Goswami, R. (2017). Big Data Security: Survey on Frameworks and Algorithms, in 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, pp. 48-54. https://doi: 10.1109/IACC.2017.0025.
[22] CNIL Open Source PIA software tool. Accessible here: https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment. Last accessed: 20 April 2020.
[23] Alnemr R. et al. (2016) A Data Protection Impact Assessment Methodology for Cloud. In: Berendt B., Engel T., Ikonomou D., Le Métayer D., Schiffner S. (eds) Privacy Technologies and Policy. APF 2015. Lecture Notes in Computer Science, vol 9484. Springer, Cham.

# A Hybrid Discrete Artificial Bee Colony for the Green Pickup and Delivery Problem with Time Windows

Djebbar Amel Mounia
University of science and technology of Oran Mohamed-Boudiaf
El Mnaouar, BP 1505, Bir El Djir 31000, Oran, Algeria
E-mail: mounia.djebbar@univ-usto.dz

Bachir Djebbar
University of science and technology of Oran Mohamed-Boudiaf
El Mnaouar, BP 1505, Bir El Djir 31000, Oran, Algeria
E-mail: bachir.djebbar@univ-usto.dz

**Keywords:** $CO_2$ emissions, combinatorial optimization, fuel consumption

**Received:** November 1, 2019

*This paper formulates a new optimization pickup and delivery problem with time windows which take into account $CO_2$ emissions. This new NP-hard combinatorial optimization problem is called green pickup and delivery problem with time windows (GPDPTW), the recent development in the vehicle routing problem and its variants, which extends PDP and PDPTW with respect to several constraints. The objective is to find a set of routes for a fleet of vehicles in order to serve given transportation requests with a minimization of fuel consumption and $CO_2$ emission to ensure the preservation of a clean and green environment. This paper presents a mathematical formulation and proposes a hybrid discrete artificial bee colony algorithm (HDABC) as a meta-heuristic algorithm which combines a discrete artificial bee colony with neighborhood operators to solve the GPDPTW model. To the best of our knowledge, this is the first time that an emission of $CO_2$ for the PDPTW is proposed. We performed computational experiments to evaluate the effectiveness of the proposed method, which provides the best result and can effectively find an optimal tour. Our results show that, (1) the shortest route is not necessarily the route that consumes the least fuel; (2) the fuel consumption is affected by the load and the number of vehicles.*

*Povzetek: Članek predstavi novo metodo za optimizacijo prevzema in dostave s časovnimi okni z minimalizacijo porabe goriva in emisij CO2.*

## 1 Introduction

Nowadays, the amount of $CO_2$ emission caused by transportation is significant for wider environmental and social impacts rather than just economic costs. It has direct effects on human health, e.g., pollution, and indirect ones, e.g., climate change. The objective of harmonizing the environmental and economic costs is to implement an effective strategy to meet the environmental concerns and financial indices. One of the most important decisions concerns the routing of vehicles with the minimum amount of $CO_2$ emissions, since it offers great potential to reduce the fuel consumption and to ensure the preservation of a clean and green environment. Thus, reducing fuel consumption can directly reduce carbon emissions. In addition, fuel consumption accounts for as much as 60% of the operating cost of a vehicle, according to [1]. Therefore, reducing fuel consumption can also reduce operating costs.

The vehicle routing problem (VRP) is a generic name given to a class of problems to determine a set of vehicle routes, in which each vehicle departs from a given depot, serves a given set of clients, and returns back to the same destination. The basic VRP involves a single depot, a fleet of identical vehicles stationed at the depot, and a set of clients who require delivery of goods from the depot. The objective of basic VRP is to minimize the total routing cost, subject to the capacity constraints on the vehicles [2] [3].

One variation of the classical vehicle routing problem considers clients that require pickup and delivery service [4]. This problem is called the Pickup and Delivery Problem with Time Windows (PDPTW). In this variant, PDPTW deals with a number of client requests that are to be served by a fleet of vehicles, while a number of constraints must be observed. Each vehicle has a limited capacity (the capacity constraint). A vehicle route usually starts and ends at a central depot. A request must be picked up from a pickup location to be delivered to a corresponding delivery location. The pickup and delivery pair must be served by the same vehicle (the coupling

constraint) and the pickup must precede the delivery (the precedence constraint). In addition, every request must be served within a predetermined time window interval (the time window constraint). If the vehicle arrives earlier than the allowed service time, it should wait until the beginning of the specified period. A vehicle may never arrive to a location after the end of the time window of the location. The PDPTW mainly involves transportation activities to complete/serve a set of requests, but the transportation has an impact on the environment due to pollution and $CO_2$ emissions.

In this paper, we consider the Green Pickups and Deliveries problem with Time Windows (GPDPTW) which is an extension of the PDPTW. The objective is to construct valid routes for the vehicles without violating vehicle capacity, time window, precedence and coupling constraints with minimal $CO_2$ emissions.

The contributions of this paper are twofold. First, the paper addresses the GPDPTW problem with precedence, coupling, capacity, time windows constraints and we introduce the factor of $CO_2$ emission, which is applied to this variant of VRP for the first time. Second, we develop and implement the Hybrid discrete artificial Bee Colony (HDABC) algorithm to solve it. To the best of our knowledge, this study is the first attempt at applying the discrete artificial bee colony meta-heuristic. Our aim is to minimize the amount of $CO_2$ emissions.

The remainder of this paper is structured as follows. Section 2 discusses some related works on Pickup and Delivery Problem with Time Windows and the Green vehicle routing problem. Section 3 focuses on the formulation of the GPDPTW. Section 4 presents a brief definition of artificial bee colony. Section 5 describes a proposed hybrid discrete artificial bee colony. Then, a case study is presented in Section 6. The last section is devoted to conclusions and the research perspectives related to the current work.

## 2 Related work

Dumas et al [5], were the first to use column generation for solving PDPTW. They proposed a branch and bound method that is able to handle problems with up to 55 requests. Sol et Savelsbergh [6] proposed a branch and price algorithm to solve the PDPTW with the objective to minimize the number of vehicles and the total travel distance. Nanry et Barnes [7] are among the first researchers to present a meta-heuristic for PDPTW. The meta-heuristic is based on a reactive tabu search. First, a feasible solution is constructed using greedy insertion method. Next, tabu search is used to improve the initial solution. Three neighborhood moves are proposed in this paper. They are: single pair insertion, swapping pairs between route and within route insertion. In order to evaluate their work, the authors created PDPTW test instances from standard vehicle routing problems with time windows proposed by Solomon [8]. Li et Lim [9] developed a hybrid meta-heuristic based on tabu search and simulated annealing to solve the PDPTW, and they also produced several test instances for the PDPTW which are generated from Solomon's 56 benchmark instances [8]. A two phase method proposed by Lau et Liang [10] was developed. In the first phase, they applied a novel construction heuristic to generate an initial solution. In the second phase, a tabu search method is proposed to improve the solution. Lim et al [11] applied "Squeaky wheel" optimization and local search to the PDPTW. Another approach to this problem was proposed by Pankratz [12], who used a grouping genetic algorithm, and this is extended to a multi-strategy grouping genetic algorithm by Ding, Li et Ju [13]. Lu et Dessouky [14] presented a new insertion-based construction heuristic to solve the multi-vehicle pickup and delivery problem with time windows. Their main contribution was to define new criteria to evaluate requests insertion based on reduction of time slack compared to the classical one based on the incremental distance measure. Bent and Van Hentenryck [15] proposed a two-stage hybrid algorithm where the first stage uses a simple simulated annealing algorithm to decrease the number of routes, while the second stage uses a large neighborhood search to decrease the total travel cost. The heuristic was tested on the problems proposed by Li et Lim [9]. In addition, Dergis et Dohmer [16] showed that the approach of indirect local search with greedy decoding gives results which are competitive with both Li et Lim [9] and Pankratz [12]. Ropke et Cordeau [17] presented a new branch and cut and price algorithm in which the lower bounds are computed by the column generation algorithm and improved by introducing different valid inequalities to the problem. More recently, ant colony System was applied by Carabetti, De Souza et Fraga [18]. Harbaoui et al [19] presented an approach based on genetic algorithms and Pareto dominance method to give a set of satisfying solutions to the PDPTW minimizing total travel cost, total tardiness time and the vehicles number.

After that, the green concept emerged as one of the latest extensions of the VRP literature in recent years. Researchers suggest that there are possibilities for reducing carbon dioxide ($CO_2$) emissions by extending the traditional VRP objectives to account for wider environmental and social impacts rather than just economic costs [20] [21] [22]. Until now, little research for minimizing energy consumption in transportation planning has been carried out, such as the PhD dissertation of Palmer [23] that presented an integrated routing and emissions model for freight vehicles and investigates the role of speed in reducing $CO_2$ emissions under various congestion scenarios and time window settings. However, Palmer did not take into account the vehicle loads in his model, although this was offered as a future research topic. Kara et al [24] considered a more realistic cost of transportation that is affected by the load of the vehicle as well as the distance of the arc travelled. They defined energy minimizing vehicle routing problem as the capacited vehicle routing problem with a new objective of cost, in which the cost function is a product of the total load (including the weight of the empty vehicle) and the length of the arc. However, they used their work to represent the energy so as to simplify the relationship between minimizing the consumed energy

and the variables of the vehicle conditions. Details of the formulation of fuel consumption are not provided. Maden et al [25] considered a vehicle routing and scheduling problem with time windows in which speed depends on the time of travel. Fagerholt et al [26] proposed an alternative solution methodology in which the arrival time was divided and the problem was solved as a shortest path problem on a directed acyclic graph. Xiao et al [27] proposed a Fuel Consumption Rate (FCR) considered Capacited Vehicle Routing Problem (CVRP), which extends CVRP with the objective of minimizing fuel consumption. In their paper, both the distance traveled and the load are considered as the factors which determine the fuel costs. FCR is taken as a load dependent function, where FCR is linearly associated with the vehicle's load. Kuo et Wang [28] developed a tabu search heuristic in order to find feasible vehicle routes while minimizing the total fuel consumption. They incorporated the effect of vehicle speed into the fuel cost. They took the travel speed as a parameter and observed the effect of this by conducting experiments on four different data sets with differing travel speed patterns. Zhang et al [29] studied the capacited vehicle routing problem from an environmental perspective and introduced a new model called environmental vehicle routing problem (EVRP) with the aim of reducing the adverse effect on the environment caused by the routing of vehicles. The environmental influence is measured through the amount of carbon dioxide emission. They designed the hybrid artificial bee colony algorithm to solve the EVRP model. Zhang et al [30] studied a vehicle routing problem (VRP) with the consideration of fuel consumption and carbon emission. They developed an improved tabu search algorithm named RS-TS for solving the model. In the RS-TS algorithm, they introduced a novel route encoding and decoding algorithm named WSS, in which three neighborhood search methods are applied. Poonthalir et Nadarajan [31] introduced a bi-objective Fuel efficient Green Vehicle Routing Problem (F-GVRP) with varying speed constraint. The problem is solved using Particle Swarm Optimization with Greedy Mutation Operator and Time varying acceleration coefficient. Liu et Jiang [32] introduced the load-dependent vehicle routing problem with time windows. They designed a new constraint relaxation-based algorithm and they presented an effective execution scheme of local search procedures. Other VRP-related studies that aim at minimizing total fuel consumption include Apaydin et Gonullu [33], Maraš [34], Nanthavanij et al [35] and Tavares et al [36].

Another problem that considers fuel consumption is the pollution routing problem (PRP). The PRP was proposed by Bektas et Laporte [22]. Its aims are to find a set of vehicle routes and vehicle speeds over the routes that minimize the operational and environmental costs, while respecting constraints on time and vehicle capacities. The PRP was addressed with a two-phase heuristic in Demir et al [37]. In the first phase, the vehicle routing problem with time windows is solved by means of an adaptive large neighborhood search, including five insertion operators and twelve removal operators. In a second phase, vehicle speeds are optimized using a recursive algorithm. A bi-objective variant considering fuel and driving time minimization is presented in Demir et al [38] and Franceschetti et al [39] considered the time-dependent PRP. Kramer et al [40] proposed a method which combines a local search-based meta-heuristic with an integer programming approach to solve the Pollution Routing Problem. This approach was also used to solve two other environmental based VRPs, namely the fuel consumption vehicle routing problem and the energy minimizing vehicle routing problem, as well as the well-known Vehicle Routing Problem with Time Windows (VRPTW) with distance minimization. Xiao et Konak [41] presented a Green Vehicle Routing and Scheduling Problem (GVRSP) considering general time-dependent traffic conditions with the primary objective of minimizing $CO_2$ emissions and weighted tardiness. They proposed a new mathematical formulation to describe the GVRSP with hierarchical objectives and weighted tardiness.

Other papers treated another variant of PDPTW that considered dynamic pickup and delivery problems with time window uncertainties [42], the same problem with time windows and electric vehicles was studied by [43] and [44] considered a setting in which a company not only has its own fleet of vehicles to service requests, but may also use the services of occasional drivers.

# 3 GPDPTW formulation

The GPDPTW can be formally defined as follows. Let G = (N, A) be a graph. The node set is N = {i ∈ N/i = 0, 1, 2,…, m}, such that m denotes a location. The node 0 denotes the depot. Since for each request we have a pair of pickup and delivery locations, the set $N^+$ = {i ∈ N / i = 1, 2,…, m/2} represents pickup locations, and the set $N^-$ = {i ∈ N / i = (m/ 2) + 1,…, m} represents delivery locations.

Each location i is associated with:

- A demand $q_i$, such that $q_i > 0$ for a pickup location, $q_i < 0$ for a delivery location and $q_i + q_j = 0$ for the same customer's pickup and delivery locations ($q_0 = 0$).
- A service time $s_i$ ($s_0 = 0$), which is the time needed to load or unload a pickup or a delivery demand.
- A time window $[e_i, l_i]$ during which the location must be served, and $l_i \geq e_i$

For each pair of nodes (i, j), travel time $t_{ij}$ and a travel distance $d_{ij}$ are specified.

The GPDPTW consists of designing a set of routes such that:

1. Each route starts and ends at the depot;
2. Each location is visited exactly once by exactly one vehicle;
3. The total vehicle load in any arc does not exceed the capacity of the vehicle assigned to it;
4. The total duration of each route (including travel and service times) does not exceed a duration limit.
5. If a vehicle arrives before the earliest pickup or delivery time of a loaction, it is allowed to wait until the start of the time window.

6. The precedence constraint requires that each pickup location must precede the corresponding delivery location.
7. The coupling constraint requires that the same pickup and delivery locations must be served by the same vehicle.

Savelsbergh et al [45] showed that the VRP is a NP-hard problem. Since the GPDPTW is a generalization of the VRP, it's a NP-hard combinatorial optimization problem, and the presence of many constraints makes the problem particularly complicated. The mathematical formulation of GPDPTW is a combination of Christofides et al [2], Savelsbergh et Sol [45], Xiao et al [27] and Zhang et al [29].

- $x_{ijk}$ a binary variable indicating whether arc (i, j) is traversed by vehicle k
- $x_{ijk} = 1$ if vehicle k traverses arc (i, j)
- $x_{ijk} = 0$ if vehicle k does not traverse arc (i, j)
- $y_{ik}$ load of vehicle k while visiting node i
- $Q_k$ capacity of vehicle k
- $D_i$ departure time from the node i / $D_i \in [e_i, l_i]$ , where $D_i = \max\{ A_i, e_i \}$
- CE the $CO_2$ emission rate
- FCR the fuel consumption rate
- $\rho_0$ the empty load FCR
- $\rho^*$ the full load FCR
- $\rho$ the FCR provided that load is q
- $q_{ijk}$ the load of vehicle while k traverses arc (i, j)

In this paper, we consider that each vehicle emits a certain amount of $CO_2$ when traveling over an arc (i, j). This amount is dependent on a number of factors, such as load, number of vehicle and distance travelled, among others. Whereas the $CO_2$ emission (CE) is fixed, it is estimated at 2.61 kg of $CO_2$ for each liter of diesel consumed [46]. The formulation of fuel consumption is provided in [27]. It is determined by both the distance traveled and the load of vehicle. Our objective is to serve all client requests while minimizing the total cost of transport. This cost is related to the $CO_2$ emission rate, the number of vehicles used and the distance travelled.

$$\textit{Minimiser } f_1 = \sum_{k\in K}\sum_{i\in N}\sum_{j\in N} X_{ijk} * d_{ijk} \quad (1)$$

$$\textit{Minimiser } f_2 = \sum_{k\in K}\sum_{i\in N}\sum_{j\in N} CE * \left(\rho_0 + \frac{\rho^* - \rho^0}{Q_k} q_{ijk}\right) * X_{ijk} * d_{ijk} \quad (2)$$

*Subject to*

$$\sum_{i=1}^{N}\sum_{k=1}^{K} x_{ijk} = 1, \qquad j = 2, \dots, N \quad (3)$$

$$\sum_{j=1}^{N}\sum_{k=1}^{K} x_{ijk} = 1, \qquad i = 2, \dots, N \quad (4)$$

$$\sum_{i=1}^{N} x_{i0k} = 1, \qquad \forall\, k \in K \quad (5)$$

$$\sum_{j=1}^{N} x_{0jk} = 1, \qquad \forall\, k \in K \quad (6)$$

$$\sum_{i=1}^{N} x_{iuk} - \sum_{j=1}^{N} x_{ujk} = 0, \forall\, k \in K, \forall\, u \in N \quad (7)$$

$$x_{ijk} = 1 \Rightarrow y_{jk} = y_{ik} + q_i, \forall k \in K, \forall\, i, j \in N \quad (8)$$

$$y_{0k} = 0, \qquad \forall k \in K \quad (9)$$

$$0 \le y_{jk} \le Q_k \qquad \forall\, k \in K, \forall\, j \in N \quad (10)$$

$$D_p \le D_d \qquad \forall\, p \in N^+, \forall\, d \in N^- \quad (11)$$

$$D_0 = 0 \quad (12)$$

$$x_{ijk} = 1 \Rightarrow D_i + t_{ijk} \le D_j, \forall k \in K, \forall\, i, j \in N \quad (13)$$

$$t_{0i} + s_i + t_{ij} < l_j, \qquad \forall\, i, j \in N, i \ne j \quad (14)$$

where $K$ is the total number of vehicles, $d_{ijk}$ is the travel distance from customer *i* to customer *j* by vehicle *k*.

Constraints (3) and (4) form the feasible routes of vehicles, so that every customer is visited by exactly one vehicle, and every vehicle that arrives to a location must leave that location. Constraints (5) and (6) ensure that each vehicle is used to serve at most one route. Constraint (7) ensures the route continuity. Constraints (8), (9) and (10) show that the total demands of any route must not exceed the capacity of the vehicle. Constraints (11), (12) and (13) ensure the precedence constraint. Constraint (14) ensures that only edges satisfying the time window constraint are allowed.

**Initialization algorithm**

```
1  Let k = 0   {k is the number of vehicles used}
2  Let list not empty {list contains all pickup
3  node}
4  repeat
5  Initialize an empty route r
6  k = k+1
7  for (All unassigned requests) do
8  A request p_i is randomly selected from list
9  Insert p_i at the end of the current route r;
10 Insert his corresponding d_i request into route r ;
11 Call the IsFeasibleSolution algorithm to
12 improve r
13 if (r is a feasible route) then
   Mark p_i as inserted
   until (All requests have been inserted)
```

# 4 Artificial bee colony

The Artificial Bee Colony (ABC) algorithm is a swarm intelligence technique inspired by the intelligent foraging behavior of honey bees. This algorithm was proposed by Karaboga et al [47] [24] [48] [49] [50] based on the foraging behaviour of honey bees. The ABC algorithm classifies the foraging artificial bees into three groups; namely, employed bees, onlookers and scouts. A bee that is currently exploiting a food source is called an employed bee. A bee waiting in the hive to make a decision in choosing a food source is named as an onlooker. A bee carrying out a random search for a new food source is called a scout. In the ABC algorithm, each solution to the problem under consideration is called a food source and represented by an n dimensional integer valued vector, whereas the fitness of the solution corresponds to the nectar amount of the associated food resource. Similar to the other swarm intelligence based approaches, the ABC algorithm is an iterative process. It starts with a population of randomly generated solutions or food sources.

The ABC algorithm is usually used for continuous optimization problems. To apply it to discrete combinatorial problems, modifications and adjustments are needed. There are several strategies to implement in each part of the algorithm, and each combination can lead to a different Discrete Artificial Bee Colony (DABC) algorithm. The point is to know which strategy and which combination among them have to be used in order to enhance the performance of the algorithm for the problem at hand [51].

# 5 Proposed HDABC for GPDPTW

The description of the proposed Hybrid Discrete Artificial Bee Colony (HDABC) is given as follows:

## 5.1 Initialization

The algorithm starts with the generation of N initial solutions. These solutions characterize the initial food sources that will be explored by the employed bees. Each food source in the discrete artificial bee colony algorithm is a feasible solution of GPDPTW, which consists of a list of routes. One route is associated with one vehicle. Each route consists of a sequence of request points (pickup and delivery) which are visited by the given vehicle. Figure 1 represents the solutions under the form of food sources where 0 represents the depot and the integer numbers represents pickup or delivery location.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | $3^+$ | $2^+$ | $2^-$ | $1^+$ | $3^-$ | $1^-$ | 0 | $4^+$ | $5^+$ | $5^-$ | $4^-$ | 0 |

Figure1: Encoding solution.

According to our method, firstly, a distributed initial population is generated. The method used for achieving initial solutions was set up in such a way that it led to achieve better quality solutions than random selection. In this paper, the initial population is created as follows; a random pickup point is selected as initial of the route, then the next requests consecutively are added to the route to ensure the constraints are satisfied and to get a feasible solution. The algorithm of initialization is defined as follows:

## 5.2 Employed bee phase

In the basic artificial bee colony algorithm, every employed bee determines a food source in the neighborhood of its currently associated food source and evaluates its nectar fitness. We know that each employed bee $x_{ij}$ generates a new food source $\acute{x}_{ij}$ in the neighborhood of its present position as follows: $\acute{x}_{ij} = x_{ij} + \varphi_{ij}(x_{ij} - x_{kj}) \qquad k = int(rand\ FN) + 1$ where $\varphi_{ij}$= (rand−0.5)×2, $\phi_{ij}$ is a uniformly distributed real random number within the range [−1, 1], FN is the number of food sources, i ∈ {1, 2,…, FN }, k ∈ {1, 2, ... , FN } and k ≠ i, j {1, 2,… , FN } are randomly chosen indexes. But this method cannot be applied to a hybrid discrete artificial bee colony. In this paper, we will propose hybrid neighborhood strategies for the HDABC algorithm to solve the GPDPTW problems. The details of the hybrid neighborhood strategies are presented in Section 5.5.

Hybrid neighborhood strategies are used to obtain a new solution $\acute{x}$ from the current solution x of the HDABC meta-heuristic. Each method for the generation of neighboring food sources may have different performance during the evolution process. The set of pre-selected operators is determined by experimental testing.

As for the selection, a new food source is always accepted if it is better than the current food source. The employed bee exploits the better solution.

## 5.3 Onlookers bee phase

After all employed bees complete their search, they come back to the hive and share their information about the nectar amount of their food sources with the onlookers waiting there; so the quality of the solutions are evaluated.

In this paper, a binary tournament is applied to choose some foods sources by onlookers. The term “binary tournament” refers to the size of two in a tournament, which is the simplest form of tournament selection [52]. Binary tournament starts by selecting two food sources at random. Then, fitness values of these food sources are evaluated. The one having more satisfactory fitness is then chosen. One advantage of the tournament selection is its ability to handle minimization problems without any structural changes. So, the onlookers play the role as an objective function to evaluate generated solutions. Obviously, when the fitness of the food source decreases, the probability with the preferred source by a looker bee decreases proportionally. The onlooker bee produces a new food source by the hybrid neighborhood strategies method presented in section 5.5, the same as the employed bee does. Then, the new source will be evaluated and compared to the primary food solution. If the new source has a better nectar amount than the primary food solution, the new source will be accepted.

## 5.4 Scout bee phase

In the standard ABC algorithm, if a solution does not improve for a predetermined number of trails ‘‘limit’’, then this food source is abandoned by its employed bee and then the employed bee becomes a scout. The scout produces a food source randomly in the search scope. But this new solution cannot carry better information for the population. In Pan et al [53] advise that the scout generates a food source by performing several insert operators to the best food source in the population. In this paper, we will use the Insertion Inter-route operator to generate a new solution.

## 5.5 Hybrid neighborhood strategies

In this paper, to enrich the neighborhood structure and diversify the population, four neighboring approaches based on the Insertion Inter-route, Swap Inert-route, Swap Intra-route and Move operator are separately

utilized to generate neighboring food sources for the employed bees and onlooker bees. On the whole, we expect that the chosen neighborhood strategies can perform distinct advantages; therefore, they can be effectively combined to solve different instances of green pickup and delivery problem with time windows. The applications of these neighborhoods are as follows:

1 : Apply Swap Intra-route to a food source.
2 : Apply Move to a food source.
3 : Apply Insertion Inter-route to a food source.
4 : Apply Swap Inter-route to a food source.

Each operator for the generation of neighboring food sources may have different performances during the evolution process. Therefore, we believe hybrid neighborhood strategies can perfectly be solvable to the green pickup and delivery problem with time windows. Based on the above considerations, we proposed a new method called Hybrid Discrete Artificial Bee Colony (HDABC), the primary idea, is that at each generation, a new bee colony is created. The detail of each operator is as follows:

**Swap Intra-route**

The role of operator swap intra-route is to improve the quality of a route by changing the order in which request points are visited. One route is selected at random. For each request from that route, we try to find a better location inside the same route. If there is such a place, we move a request to that place which satisfies all constraints for the problem. The swap intra-route operator is shown in Figure 2 (see Appendix).

**Move**

The role of move operator is to find the best position by changing the order in which request points (pickup-delivery) are visited. For each request from that route we move a location inside or beside that route. If there is such a place, we move a request to that place which satisfies all constraints for the problem. The move operator is shown in Figure 3 (see Appendix).

**Insertion Inter-route**

The insertion inter-route moves a pickup-delivery pair from its current route to another route in the solution. They perform the following process for all pickup-delivery pairs in the current solution. An admissible placement is one where both requests (pickup and delivery) satisfy all the constraints of the problem. To reduce the number of routes, the search process should be biased such that it tries to remove the request pairs from the shorter routes and insert them into longer routes which satisfy all constraints for the problem. The insertion inter-route operator is shown in Figure 4 (see Appendix).

**Swap Inter-route**

Swapping randomly requested pairs, i.e., a pickup followed by a delivery node between two different routes. For each pair, we check whether it can be relocated by exchanging its pickup and delivery positions with the pickup and delivery positions of any other request pair in another route which satisfies all constraints for the problem. The swap inter-route operator is shown in figure 5 (see Appendix).

### 5.6 Proposed HDABC

In this paper, we propose a new method called Hybrid Discrete Artificial Bee Colony (HDABC), the primary idea is to hybridize neighborhood strategies at each generation to create a new bee colony. The above idea is illustrated in figure 6.

In the proposed HDABC, there are four strategies to update the food sources. We present a food source as a route and apply the discrete operations to generate new neighborhood food source for three different bees. The heuristic in section 5.1 was used to initialize the population with certain quality and diversity. Then, the new hybrid neighborhood strategies were used to solve the green pickup and delivery problem with time windows. The procedure of HDABC proposed, is given as follows:

## 6 Case study

### 6.1 Data and parameters setting

A numerical example is used to illustrate the applications of the proposed model and the solution algorithm. The data sets for the problem are derived from the instances created by Li et Lim [9] which are related to the well known Solomon instances. The datasets are available at the following link:
http://www.sintef.no/Projectweb/TOP/PDPTW.

The graph consists of one depot and 40 nodes. In each instance, nodes are located in geographical clusters and have a small vehicle capacity and narrow time windows. Instances were generated considering only the first 40 nodes which are represented by a complete graph, and all distances are Euclidean distances satisfying the triangle inequality. The pickup and delivery requests are paired and therefore, the number of nodes in the network, without the vehicle depots, is even. Each node has x and y coordinates, a demand $q_i$, a time window $[e_i, l_i]$, a service time $s_i$ and the corresponding pickup (succ) or delivery (pred) node. The following table (Table 1) represents an example of instance used for the problem. The $CO_2$ emission rate (CE) per liter of fuel, the fuel consumption rate for both empty-load ($\rho_0$) and full-load ($\rho^*$) situations are set to 2.61, empty load fuel consumption rate = 0.296 and full load fuel consumption rate 0.390, respectively referring to a previous case study by [46].

**HDABC Algorithm:**

**Set** Popsize = FN //Colony size = 2*FN
**Set** Max.iter = Maximum number of iterations
**Set** Max.trial = Maximum number of improvement trials
$E_i = \emptyset$; $i = \langle 1,\ldots..,n \rangle$, $E_i$ is the set of neighbor solution of food source
Generate FN food sources using the method presented in section 5.1 for initial population
Evaluate initial population // Calculate $f(x_i)$ for each food sources
Memorize best food source $x_i$ ;
**Set** iteration = 1
**For each** food source i **do, Set** $trial_i$ = **0 end for**
**do while** iteration $\leq$ Max.iter
//*****EMPLOYED BEE PHASE*****
**For** each food source $x_i$ **do**
Apply hybrid neighborhood strategies //Produce a new neighbor solution $\acute{x}_i$
Evaluate $\acute{x}_i$ //Calculate $f(\acute{x}_i)$
**If (** $f(x_i) > f(\acute{x}_i)$ **) then**
replace $x_i$ with $\acute{x}_i$
**Set** $trial_i = 0$
**Else**
**Set** $trial_i = trial_i + 1$
**EndIf**
**End For**
//*****ONLOOKER BEE PHASE*****
**For** each onlooker **do**
Select a food source using the binary tournament selection method
Apply hybrid neighborhood strategies //Produce a new neighbor solution $\acute{x}_i$
$E_i = E_i \cup \acute{x}_i$
**End For**
**For** each food source $x_i$ and $E_i \neq \emptyset$ **do**
**If (** $f(x_i) > f(\acute{x}_i)$ **) then**
replace $x_i$ with $\acute{x}_i$
**Set** $trial_i = 0$
**Else**
**Set** $trial_i = trial_i + 1$
**End If**
**End For**
//*****SCOUT BEE PHASE*****
**Set** i = index of max(trial) //Find the index that has the maximum trial value
**If (**$trial_i$ >=Max.trial **) then**
replace $x_i$ with Insertion Inter-route operator
**End If**
Memorize global best solution
**end while**

A robust parameter setting is required for the proposed HDABC algorithm to efficiently perform on different data sets. In order to select best parameter setting, tests are performed on three parameters: FN (Population of food sources), number of iterations, and limit (number of trails). Since an obvious correlation between the optimal settings of these parameters is observed, each one of them is tested individually for deciding on the standard parameter setting. After several preliminary experiments, the size of population is fixed to 100, the number of iterations is fixed as 200 and the limit is 20.

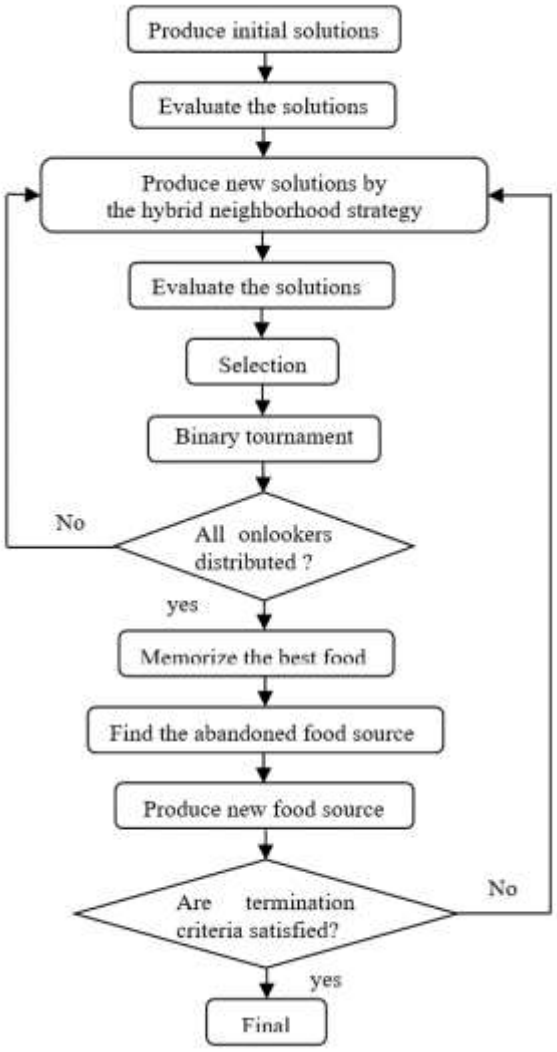



Figure 6:The flow chart of HDABC algorithm.

## 6.2 Analysis and discussion of results

In this section, we present a numerical example for the GPDPTW problem and the corresponding optimum solution that is obtained from the proposed HDABC algorithm described in Section 5 which is coded in C++ Builder 2010 software running on a personal computer using Intel Core i5, 2.60 gigahertz, 64-bits processor with 4 gigabyte RAM and Windows 8 OS. In order to analyze the performance of HDABC which is applying from the first time to the green pickup and delivery problem with time windows, we use problem instances with 40 nodes.

| Request | x | y | q | e | l | S | succ | Pred |
|---|---|---|---|---|---|---|---|---|
| 0 | 35 | 35 | 0 | 0 | 110 | 0 | 0 | 0 |
| 1 | 41 | 49 | 10 | 55 | 88 | 10 | 0 | 5 |
| 2 | 35 | 17 | 7 | 59 | 89 | 10 | 0 | 20 |
| 3 | 55 | 45 | -23 | 30 | 87 | 10 | 18 | 0 |
| …. | | | | | | | | |

Table 1: An example instance.

For proper comparison of the computational results, we conduct our experiments upon two cases; case one with the objective of the minimum amount of $CO_2$ emission and case two with the objective of shortest distance. We run the GPDPTW 10 times, and we compare the average minimum emission of $CO_2$ and the average distance obtained for the two cases during the ten runs. The results are tabulated in tables 2–4 (see Appendix).

All 40 nodes problem instances are solved to optimality. For our approach, we report a column $CO_2$ gap (%), which is the relative gap between the amount of $CO_2$ found in the case of minimum $CO_2$ emission and the amount of $CO_2$ found in the case of shortest distance. For example, assuming that the amount of emissions of $CO_2$ for an instance in case 1 is $CO_{2min}$ and the amount of emissions of $CO_2$ for the same instance in case 2 is $CO_{2max}$, then gap (%) is calculated as $100 \times (CO_{2min} / CO_{2max} - 1)$ and a column distance gap(%), which is the relative gap between the total travelled distance found in the case of minimum $CO_2$ emission and the total travelled distance found in the case of shortest distance. For example, assuming that the total travelled distance for an instance in case 1 is $dist_{max}$ and the total travelled distance for the same instance in case 2 is $dist_{min}$, then gap(%) is calculated as $100 \times (dist_{max} / dist_{min} - 1)$. We can notice that comparing the case of minimum $CO_2$ emission and the case of the shortest distance, the amount of $CO_2$ could be reduced by 3,43% on average and the average on total distance traveling is increased by 4,60% , we can deduce that the amount of $CO_2$ emission is not guaranteed for the vehicles along the shortest path. Thus, the distances between customers are shorter and the loading rate and the number of vehicles used in the routes are higher.

Figure 7 shows that lc101, lc104, lc105, lc106 and lc109 have significant effects on the amount of $CO_2$ emissions. The savings percentages are respectively -5,52%, 3,67%, 9,81%, 3,36% and 4,68%. In these instances, the second case, in addition to the total load of the vehicle, the number of vehicles used is reduced compared to the first case. Thus, the fuel consumption is reduced.

# 7 Conclusion

This paper proposes and develops a hybrid discrete artificial bee colony approach to solve and discuss the green pickup and delivery problem with time windows (GPDPTW), the recent development in the vehicle routing problem and its variants, which extends the classical vehicle routing problem by considering the coupling, the precedence, the time windows constraints

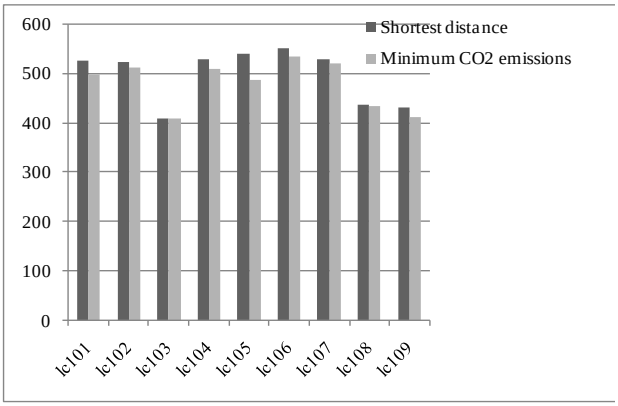



Figure 7: The difference between the amounts of CO2 emissions in the two cases.

and the $CO_2$ emission by vehicles which make the NP-hard combinatorial and optimization problem. The major contribution of this paper is two-fold. First, the GPDPTW model is presented and formulated. Second, a hybrid discrete artificial bee colony for solving the HDABC model is developed which combines a discrete artificial bee colony meta-heuristic with neighborhood operators. The objective is to minimize the amount of fuel consumption to minimize the $CO_2$ emissions while respecting all constraints. Costs are based on fuel consumption which depends on many factors, such as travel distance, vehicle load and the number of vehicles. The solution approach is evaluated in terms of optimality to reach the best solution on the various test instances. Computational experiments show that the proposed method is effective and efficient and can solve the problem optimally.

Research perspectives in the field highlight the application of the proposed method to accommodate multiple depot and heterogeneous vehicles. The GPDPTW is formulated by assuming only one depot and homogeneous vehicles. However, in many cases, companies may have multiple depots and different kinds of vehicle for pickup and delivery operations.

# 8 References

[1] B. Sahin, H. Yilmaz, Y. Ust, A. F. Guneri, et B. Gulsun, « An approach for analysing transportation costs and a case study », *European Journal of Operational Research*, vol. 193, nº 1, p. 1-11, févr. 2009. https://doi.org/10.1016/j.ejor.2007.10.030.

[2] N. Christofides, A. Mingozzi, et P. Toth, «*The vehicle routing problem* ». Chichester; New York: Wiley, 1979.

[3] P. Toth et D. Vigo, « The Vehicle Routing Problem », Édition Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2002, p. 1–26. https://doi.org/10.1137/1.9780898718515.

[4] H. Xu, Z.-L. Chen, S. Rajagopal, et S. Arunapuram, « Solving a Practical Pickup and Delivery Problem », *Transportation Science*, vol. 37, nº 3, p. 347-364, août 2003. https://doi.org/10.1287/trsc.37.3.347.16044.

[5] Y. Dumas, J. Desrosiers, et F. Soumis, « The pickup and delivery problem with time windows », *European Journal of Operational Research*, vol. 54, nº1, p. 7-22, sept. 1991. https://doi.org/10.1016/0377-2217(91)90319-Q.

[6] M. Sol et M. W. P. Savelsbergh, « A branch-and-price algorithm for the pickup and delivery problem with time windows », *Memorandum COSOR*, vol. 9422, 1994.

[7] W. P. Nanry et J. Wesley Barnes, « Solving the pickup and delivery problem with time windows using reactive tabu search », *Transportation Research Part B: Methodological*, vol. 34, nº 2, p. 107-121, févr. 2000. https://doi.org/10.1016/S0191-2615(99)00016-8.

[8] M. M. Solomon, « Algorithms for the Vehicle Routing and Scheduling Problems with Time Window Constraints », *Operations Research*, vol. 35, nº 2, p. 254-265, avr. 1987. https://doi.org/10.1287/opre.35.2.254.

[9] H. Li et A. Lim, « A metaheuristic for the pickup and delivery problem with time windows », in *Proceedings 13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2001*, nov. 2001, p. 160-167. https://doi.org/10.1109/ICTAI.2001.974461.

[10] H. C. Lau et Z. Liang, « Pickup and delivery with time windows: algorithms and test case generation », in *Proceedings 13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2001*, nov. 2001, p. 333-340. https://doi.org/10.1109/ICTAI.2001.974481.

[11] H. Lim, A. Lim, et B. Rodrigues, « Solving the pickup and delivery problem with time windows using"squeaky wheel" optimization with local search », 2002, p. 2335-2344.

[12] G. Pankratz, « A Grouping Genetic Algorithm for the Pickup and Delivery Problem with Time Windows », *OR Spectrum*, vol. 27, nº 1, p. 21-41, janv. 2005. https://doi.org/10.1007/s00291-004-0173-7.

[13] G. Ding, L. Li, et Y. Ju, « Multi-strategy grouping genetic algorithm for the pickup and delivery problem with time windows », in *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation - GEC '09*, Shanghai, China, 2009, p. 97. https://doi.org/10.1145/1543834.1543849.

[14] Q. Lu et M. M. Dessouky, « A new insertion-based construction heuristic for solving the pickup and delivery problem with time windows », *European Journal of Operational Research*, vol. 175, nº 2, p. 672-687, déc. 2006. https://doi.org/10.1016/j.ejor.2005.05.012.

[15] R. Bent et P. V. Hentenryck, « A two-stage hybrid algorithm for pickup and delivery vehicle routing problems with time windows », *Computers & Operations Research*, vol. 33, nº 4, p. 875-893, avr. 2006. https://doi.org/10.1016/j.cor.2004.08.001.

[16] U. Derigs et T. Döhmer, « Indirect search for the vehicle routing problem with pickup and delivery and time windows », *OR Spectrum*, vol. 30, nº 1, p. 149-165, janv. 2008. https://doi.org/10.1007/s00291-006-0072-1.

[17] S. Ropke et J.-F. Cordeau, « Branch and Cut and Price for the Pickup and Delivery Problem with Time Windows », *Transportation Science*, vol. 43, nº 3, p. 267-286, juin 2009. https://doi.org/10.1287/trsc.1090.0272.

[18] E. G. Carabetti, S. R. d Souza, M. C. P. Fraga, et P. H. A. Gama, « An Application of the Ant Colony System Metaheuristic to the Vehicle Routing Problem with Pickup and Delivery and Time Windows », in *2010 Eleventh Brazilian Symposium on Neural Networks*, oct. 2010, p. 176-181. https://doi.org/10.1109/SBRN.2010.38.

[19] I. Harbaoui Dridi, R. Kammarti, M. Ksouri, et P. Borne, « Genetic Algorithm for Mulicriteria Optimization of a Multi-Pickup and Delivery Problem with Time Windows », in *INCOM'09 IFAC*, Russia, juin 2009, p. 1521-1526. https://doi.org/10.12700/APH.12.8.2015.8.9.

[20] A. C. McKinnon et M. I. Piecyk, « Measurement of CO2 emissions from road freight transport: A review of UK experience », *Energy Policy*, vol. 37, nº 10, p. 3733-3742, oct. 2009. https://doi.org/10.1016/j.enpol.2009.07.007.

[21] A. Sbihi et R. W. Eglese, « Combinatorial optimization and Green Logistics », *4OR*, vol. 5, nº 2, p. 99-116, juill. 2007. https://doi.org/10.1007/s10288-007-0047-3.

[22] T. Bektaş et G. Lapory a dete, « The Pollution-Routing Problem », *Transportation Research Part B: Methodological*, vol. 45, nº 8, p. 1232-1250, sept. 2011. https://doi.org/10.1016/j.trb.2011.02.004.

[23] A. Palmer, « The development of an integrated routing and carbon dioxide emissions model for goods vehicles », *Ph.D. Dissertation*, School of Management, Cranfield University, nov. 2007.

[24] İ. Kara, B. Y. Kara, et M. K. Yetis, « Energy Minimizing Vehicle Routing Problem », in *Combinatorial Optimization and Applications*, 2007, p. 62-71. https://doi.org/10.1007/978-3-540-73556-4_9.

[25] W. Maden, R. Eglese, et D. Black, « Vehicle routing and scheduling with time-varying data: A case study », *Journal of the Operational Research Society*, vol. 61, nº 3, p. 515-522, mars 2010. https://doi.org/10.1057/jors.2009.116.

[26] K. Fagerholt, « Optimal fleet design in a ship routing problem », *International Transactions in Operational Research*, vol. 6, nº 5, p. 453-464, 1999. https://doi.org/10.1111/j.1475-3995.1999.tb00167.x.

[27] Y. Xiao, Q. Zhao, I. Kaku, et Y. Xu, « Development of a fuel consumption optimization model for the capacitated vehicle routing problem », *Computers & Operations Research*, vol. 39, nº 7, p. 1419-1431, juill. 2012. https://doi.org/10.1016/j.cor.2011.08.013.

[28] Y. Kuo et C.-C. Wang, « Optimizing the VRP by minimizing fuel consumption », *Management of Environmental Quality: An International Journal*, juin 2011. https://doi.org/10.1108/14777831111136054.

[29] S. Zhang, C. K. M. Lee, K. L. Choy, W. Ho, et W. H. Ip, « Design and development of a hybrid artificial bee colony algorithm for the environmental vehicle routing problem », *Transportation Research Part D: Transport and Environment*, vol. 31, p. 85-99, août 2014. https://doi.org/10.1016/j.trd.2014.05.015.

[30] J. Zhang, Y. Zhao, W. Xue, et J. Li, « Vehicle routing problem with fuel consumption and carbon emission », *International Journal of Production Economics*, vol. 170, p. 234-242, déc. 2015. https://doi.org/10.1016/j.ijpe.2015.09.031.

[31] G. Poonthalir et R. Nadarajan, « A Fuel Efficient Green Vehicle Routing Problem with Varying Speed Constraint (F-GVRP) », *Expert Syst. Appl.*, vol. 100, nº C, p. 131–144, juin 2018. https://doi.org/10.1016/j.eswa.2018.01.052.

[32] R. Liu et Z. Jiang, « A constraint relaxation-based algorithm for the load-dependent vehicle routing problem with time windows », *Flexible Services and Manufacturing Journal*, vol. 31, nº 2, p. 331-353, 2019. https://doi.org/10.1007/s10696-018-9323-0.

[33] O. Apaydin et M. T. Gonullu, « Emission control with route optimization in solid waste collection process: A case study », *Sadhana*, vol. 33, nº 2, p. 71-82, avr. 2008. https://doi.org/10.1007/s12046-008-0007-4.

[34] V. Maraš, « Determining Optimal Transport Routes of Inland Waterway Container Ships », *Transportation Research Record*, vol. 2062, nº 1, p. 50-58, janv. 2008. https://doi.org/10.3141/2062-07.

[35] S. Nanthavanij, P. Boonprasurt, W. Jaruphongsa, et V. Ammarapala, « Vehicle Routing Problem with Manual Materials Handling: flexible delivery crew-vehicle assignments », Bali, Indonesia, 2008.

[36] G. Tavares, Z. Zsigraiova, V. Semiao, et M. da G. Carvalho, « A case study of fuel savings through optimisation of MSW transportation routes », *Management of Environmental Quality: An International Journal*, juin 2008. https://doi.org/10.1108/14777830810878632.

[37] E. Demir, T. Bektaş, et G. Laporte, « An adaptive large neighborhood search heuristic for the Pollution-Routing Problem », *European Journal of Operational Research*, vol. 223, nº 2, p. 346-359, déc. 2012. https://doi.org/10.1016/j.ejor.2012.06.044.

[38] E. Demir, T. Bektaş, et G. Laporte, « The bi-objective Pollution-Routing Problem », *European Journal of Operational Research*, vol. 232, nº 3, p. 464-478, févr. 2014. https://doi.org/10.1016/j.ejor.2013.08.002.

[39] A. Franceschetti, D. Honhon, T. Van Woensel, T. Bektaş, et G. Laporte, « The time-dependent pollution-routing problem », *Transportation Research Part B: Methodological*, vol. 56, p. 265-293, oct. 2013. https://doi.org/10.1016/j.trb.2013.08.008.

[40] R. Kramer, A. Subramanian, T. Vidal, et L. dos A. F. Cabral, « A matheuristic approach for the Pollution-Routing Problem », *European Journal of Operational Research*, vol. 243, nº 2, p. 523-539, juin 2015. https://doi.org/10.1016/j.ejor.2014.12.009.

[41] Y. Xiao et A. Konak, « A simulating annealing algorithm to solve the green vehicle routing & scheduling problem with hierarchical objectives and weighted tardiness », *Applied Soft Computing*, vol. 34, p. 372-388, sept. 2015. https://doi.org/10.1016/j.asoc.2015.04.054.

[42] P. Györgyi et T. Kis, « A probabilistic approach to pickup and delivery problems with time window uncertainty », *European Journal of Operational Research*, vol. 274, nº 3, p. 909-923, mai 2019. https://doi.org/10.1016/j.ejor.2018.10.031.

[43] D. Goeke, « Granular tabu search for the pickup and delivery problem with time windows and electric vehicles », *European Journal of Operational Research*, vol. 278, nº 3, p. 821-836, nov. 2019. https://doi.org/10.1016/j.ejor.2019.05.010.

[44] L. Dahle, H. Andersson, M. Christiansen, et M. G. Speranza, « The pickup and delivery problem with time windows and occasional drivers », *Computers & Operations Research*, vol. 109, p. 122-133, sept. 2019. https://doi.org/10.1016/j.cor.2019.04.023.

[45] M. W. P. Savelsbergh et M. Sol, « The General Pickup and Delivery Problem », *Transportation Science*, vol. 29, nº 1, p. 17-29, 1995. https://doi.org/10.1287/trsc.29.1.17

[46] S. Ubeda, F. J. Arcelus, et J. Faulin, « Green logistics at Eroski: A case study », *International Journal of Production Economics*, vol. 131, nº 1, p. 44-51, mai 2011. https://doi.org/10.1016/j.ijpe.2010.04.041.

[47] D. Karaboga, « An idea based on honey bee swarm for numerical optimization », *Technical Report-TR06*, Department of Computer Engineering, Erciyes University, 2005.

[48] D. Karaboga et B. Basturk, « On the performance of artificial bee colony (ABC) algorithm », *Applied Soft Computing*, vol. 8, nº 1, p. 687-697, janv. 2008. https://doi.org/10.1016/j.asoc.2007.05.007.

[49] D. Karaboga et B. Akay, « A comparative study of Artificial Bee Colony algorithm », *Applied Mathematics and Computation*, vol. 214, nº 1, p. 108-132, août 2009. https://doi.org/10.1016/j.amc.2009.03.090.

[50] N. Karaboga, « A new design method based on artificial bee colony algorithm for digital IIR filters », *Journal of the Franklin Institute*, vol. 346, nº 4, p. 328-348, mai 2009. https://doi.org/10.1016/j.jfranklin.2008.11.003.

[51] W. Y. Szeto, Y. Wu, et S. C. Ho, « An artificial bee colony algorithm for the capacitated vehicle routing problem », *European Journal of Operational Research*, vol. 215, nº 1, p. 126-135, nov. 2011. https://doi.org/10.1016/j.ejor.2011.06.006.

[52] [52] A. Brindle, « Genetic algorithms for function optimization », *Doctoral dissertation*, University of Alberta, Edmonton, Canada, 1980.

[53] Q.-K. Pan, L. Wang, J.-Q. Li, et J.-H. Duan, « A novel discrete artificial bee colony algorithm for the hybrid flowshop scheduling problem with makespan minimisation », *Omega*, vol. 45, p. 42-56, juin 2014. https://doi.org/10.1016/j.omega.2013.12.004.

## Appendix

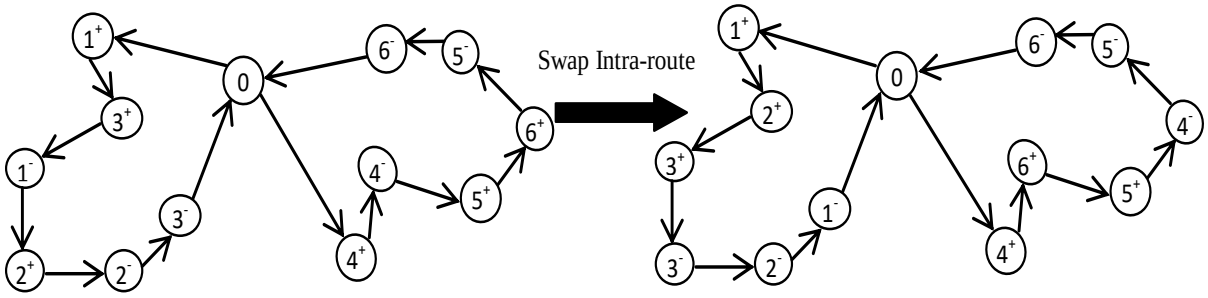



Figure 2: Swap Intra-route Operator.

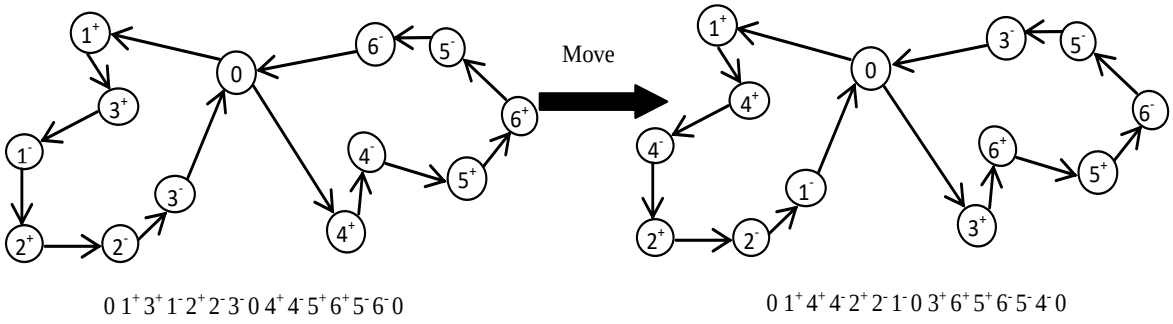



Figure 3: Move Operator.

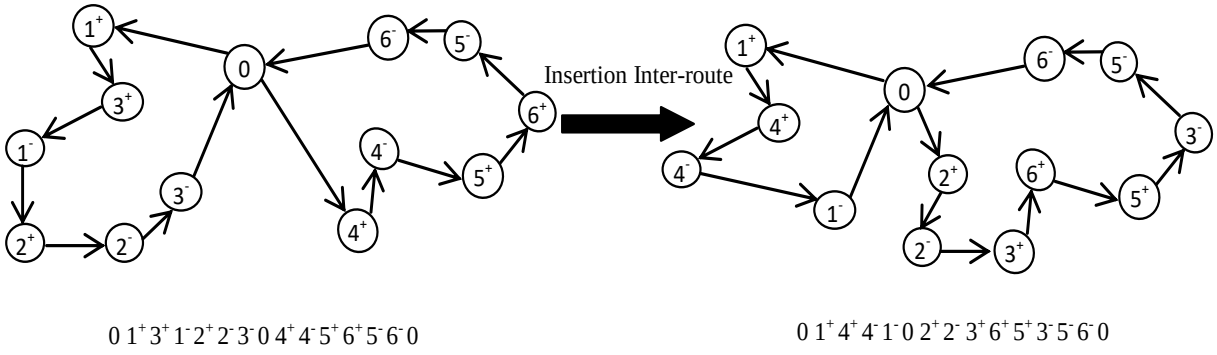



Figure 4: Insertion Inter-route Operator.

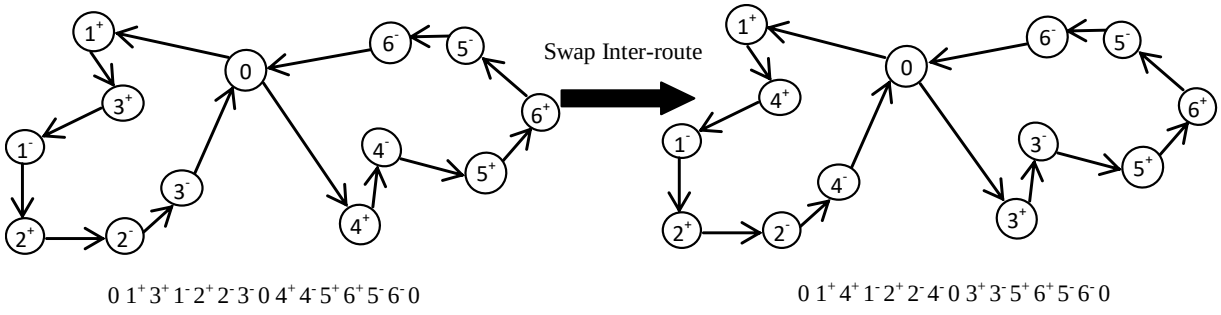



Figure 5: Swap Inter-route Operator.

| Instance | Case 1-Minimum $CO_2$ | | | Case 2-Shortest distance | | | $CO_2$ gap(%) | distance gap(%) |
|---|---|---|---|---|---|---|---|---|
| | $CO_2$ | distance | vehicle | $CO_2$ | Distance | Vehicle | | |
| lc101 | 498,31 | 686,63 | 4 | 527,43 | 643,49 | 5 | -5,52 | 6,7 |
| lc102 | 513,39 | 675,17 | 3 | 523,60 | 668,29 | 3 | -1,95 | 1,03 |
| lc103 | 410,91 | 532,75 | 2 | 411,01 | 523,14 | 2 | -0,02 | 1,84 |
| lc104 | 510,30 | 690,23 | 3 | 529,74 | 636,21 | 4 | -3,67 | 8,49 |
| lc105 | 487,41 | 699,78 | 4 | 540,43 | 627,90 | 5 | -9,81 | 11,45 |
| lc106 | 535,09 | 717,36 | 4 | 553,69 | 694,60 | 5 | -3,36 | 3,28 |
| lc107 | 522,01 | 690,23 | 4 | 529,74 | 675,09 | 4 | -1,46 | 2,24 |
| lc108 | 436,10 | 570,26 | 6 | 437,74 | 559,07 | 6 | -0,37 | 2 |
| lc109 | 413,46 | 563 | 4 | 433,77 | 539,64 | 5 | -4,68 | 4,33 |
| Avg | 480,78 | 647,27 | - | 498,57 | 618,60 | - | -3,43 | 4,6 |

Table 2: Comparative analysis between the shortest distance and the minimum CO2 emissions.

| Instance | Distance | vehicle | $CO_2$ emissions | Visited nodes |
|---|---|---|---|---|
| lc101 | 643,49 | 5 | 527,34 | 0 20 24 32 31 18 19 8 10 15 16 14 12 6 2 0<br>0 33 37 30 11 28 9 4 22 1 21 0<br>0 3 13 17 35 39 23 0<br>0 38 34 0<br>0 5 7 25 27 40 29 26 36 0 |
| lc102 | 668,29 | 3 | 523,60 | 0 24 32 33 25 27 40 29 26 1 7 11 30 38 36 37 34 23 21 0<br>0 13 18 31 35 8 2 3 10 15 12 5 28 39 9 6 4 0<br>0 20 17 19 16 14 22 0 |
| lc103 | 532,14 | 2 | 411,01 | 0 32 33 18 17 19 16 25 40 35 27 30 31 38 37 39 36 34 23 4 2 1 26 29 28 22 21 20 24 0<br>0 15 14 13 3 8 12 10 5 11 7 9 6 0 |
| lc104 | 636,21 | 4 | 529,74 | 0 17 19 40 35 5 15 14 11 31 38 0<br>0 37 39 36 34 4 2 36 34 23 4 2 1 0<br>0 26 29 28 22 21 20 24 0<br>0 15 14 13 3 8 12 10 5 11 7 9 6 0 |
| lc105 | 627,90 | 5 | 540,43 | 0 5 3 19 15 37 39 26 28 22 21 0<br>0 7 29 30 9 16 14 23 6 2 1 0<br>0 25 27 40 10 11 34 0<br>0 20 24 17 13 18 32 33 31 35 38 36 12 0<br>0 8 4 0 |
| lc106 | 694,60 | 5 | 553,69 | 0 20 25 27 29 30 38 39 28 9 2 0<br>0 3 31 40 8 4 36 34 22 0<br>0 24 18 19 15 14 23 0<br>0 7 5 13 17 33 32 35 37 10 11 0<br>0 16 12 26 6 1 21 0 |
| lc107 | 675,09 | 4 | 529,74 | 0 5 7 3 10 11 9 16 12 28 21 0<br>0 40 8 2 23 0<br>0 33 36 0<br>0 20 24 32 31 25 13 17 18 19 15 27 35 37 29 14 30 39 38 34 6 26 22 4 1 0 |
| lc108 | 559,07 | 6 | 437,74 | 0 24 31 35 29 38 36 0<br>0 20 32 40 15 12 39 34 22 0<br>0 13 17 18 19 16 14 26 23 0<br>0 25 33 37 30 28 21 0<br>0 27 2 0<br>0 5 3 7 8 10 11 9 6 4 1 0 |
| lc109 | 539,64 | 5 | 433,77 | 0 32 33 31 37 38 34 0<br>0 20 18 15 22 0<br>0 29 26 13 12 24 25 40 27 0<br>0 8 3 10 2 30 7 11 5 9 28 0<br>0 39 37 34 6 4 23 21 1 0 |

Table 3: Results concerning shortest distance.

| **Instance** | **Distance** | **vehicle** | **$CO_2$ emissions** | **Visited nodes** |
|---|---|---|---|---|
| lc101 | 686,63 | 4 | 498,31 | 0 18 29 26 12 6 2 0<br>0 3 13 17 33 37 38 34 23 0<br>0 30 28 22 21 5 7 40 35 39 36 0<br>0 20 24 32 31 25 27 19 8 10 15 11 16 14 9 4 1 0 |
| lc102 | 675,17 | 3 | 513,39 | 0 31 32 33 17 19 35 38 16 14 36 0<br>0 20 18 15 22 0<br>0 29 26 13 12 24 25 40 27 8 3 10 2 30 7 11 5 9 28 39 37 34 6 4 23 21 1 0 |
| lc103 | 532,75 | 2 | 410,91 | 0 32 33 17 18 16 19 25 40 35 27 30 31 38 37 39 36 34 23 4 2 22 21 1 26 29 28 20 24 0<br>0 15 14 13 3 8 12 10 5 11 7 9 6 0 |
| lc104 | 690,23 | 3 | 510,30 | 0 17 32 31 18 13 19 35 37 27 29 15 14 16 12 6 4 26 22 0<br>0 5 3 7 20 30 11 10 9 28 38 34 21 0<br>0 8 40 23 2 24 25 33 36 39 1 0 |
| lc105 | 699,78 | 4 | 487,41 | 0 20 24 17 25 31 35 27 19 29 30 15 16 14 12 23 6 2 1 0<br>0 40 28 22 34 0<br>0 32 33 38 36 0<br>0 5 3 13 18 8 7 10 11 37 39 9 4 26 21 0 |
| lc106 | 717,36 | 4 | 535,09 | 0 33 37 26 21 18 19 15 14 6 1 0<br>0 20 25 7 11 9 16 12 2 0<br>0 24 32 35 23 8 4 5 10 0<br>0 3 13 17 31 40 27 29 30 38 39 28 36 34 22 0 |
| lc107 | 690,23 | 4 | 522,01 | 0 17 32 31 18 13 19 35 37 27 29 15 14 16 12 6 4 26 22 0<br>0 5 3 7 20 30 11 10 9 28 38 34 21 0<br>0 8 40 23 2 0<br>0 24 25 33 36 39 1 0 |
| lc108 | 570,26 | 6 | 436,10 | 0 33 31 35 37 11 9 6 4 0<br>0 27 40 39 38 36 2 0<br>0 3 24 29 10 28 21 0<br>0 25 7 8 30 0<br>0 5 32 15 12 34 1 0<br>0 20 13 17 18 19 16 14 26 23 22 0 |
| lc109 | 563 | 4 | 413,46 | 0 20 24 8 10 29 26 9 2 23 21 7 19 12 4 0<br>0 25 5 3 30 27 11 6 1 32 37 0<br>0 33 31 40 35 39 36 38 34 0<br>0 13 17 18 16 15 14 28 22 0 |

Table 4: Results with minimum of CO2 emissions.

# Research on Recognition and Classification of Folk Music Based on Feature Extraction Algorithm

Xi Wang
Henan Polytechnic, Zhengzhou, Henan 450046, China
E-mail: xi33n9@yeah.net

**Keywords:** folk music, feature extraction, music classification, support vector machine

**Received:** December 8, 2020

*In this study, the feature extraction algorithm for folk music was analyzed. The features of folk music were extracted in aspects of time domain and frequency domain. Then, a support vector machine (SVM) was selected to identify and classify folk music. It was found that the performance of SVM was the best when $\sigma^2$ was 26 and $C$ was 4; the recognition rate of using only one feature was inferior to that of using all features; the highest recognition rate of SVM was 92.76%; compared with back propagation neural network (BPNN) and decision tree classification method, SVM had a higher recognition rate. The experimental results show the effectiveness of SVM, which can be applied in practice.*

*Povzetek: V tem študentskem članku je predstavljena klasifikacija glasbe s pomočjo metod umetne inteligence.*

# 1 Introduction

As an art form, music can express people's thoughts, feelings, and life style and has a role in promoting people's emotion and spirit. With the improvement of human living standards, music has become more and more popular. With the development of science and technology, more and more people have tended to enjoy music through the Internet. Therefore, finding out music which users want to listen to from a massive amount of music has become more and more important, and the recognition and classification of music have attracted more and more extensive attention. Huang et al. [1] improved the hidden Markov model (HMM) using an artificial neural network (ANN). The application of the improved HMM in practical music classification found that HMM had a fast calculation speed but a poor classification performance, ANN had a good classification performance but a high computational complexity. The combination of them could improve the recognition rate of HMM by 4% - 5% while maintaining the same calculation speed as HMM. Abidin et al. [2] recognized a Turkish music data set, SymbTr, with ten machine learning algorithms, and found that the performance of the algorithms was between 82% and 88%. Rao et al. [3] studied chord recognition. Pitch Class Profile features were extracted from raw audio and recognized by spare representation. Through the experiment on MIREX09, it was found that the method had robustness to Gaussian white noise. Iloga et al. [4] studied the genre classification of music, designed a sequential pattern mining method, and carried out experiments on GTZAN. They found that the accuracy of the method was 91.6%, which was more than 7% higher than the existing classifiers. Chinese folk music refers to the music played by traditional instruments, which has high artistry and nationality [5], but there is little research on its recognition and classification. Therefore, this study took folk music as the research subject, carried out feature extraction in aspects of time domain and frequency domain, established a feature database, and then identified and classified folk music with a support vector machine (SVM), and verified the reliability of the method through experiments. The present study contributes to the realization of the automatic classification of folk music and the improvement of retrieval efficiency.

# 2 Folk music and feature extraction

## 2.1 Folk music

Folk music includes instrumental music, songs, opera, etc. Musical instruments play a very important role in folk music, which can be divided into four categories, as shown in Table 1.

| | |
|---|---|
| Wind instruments | Xiao, Suona, Lusheng, Xun, pan flute, etc. |
| Plucked stringed instrument | Chinese lute, moon lute, Guqin, kayagum, Zheng, Konghou, etc. |
| Percussion instruments | Collected bronze bells, wooden fish, bronze drum, long drum, gong, etc. |
| String instruments | Erhu, Xiqin, horse head string instrument, Leiqin, etc. |

Table 1: Folk music instruments.

Musical instruments can be solo or ensemble, and different combinations of musical instruments will form different styles of instrumental music. For example, the music played by percussion instruments has a strong

rhythm and rich timbre; the music performed with string instruments has a delicate style and simple and elegant style; the music played with wind instruments, and string instruments tends to be light and lively; the music played with wind instruments and percussion instruments is joyful and enthusiastic.

## 2.2 Music feature extraction

To identify and classify folk music, it is necessary to extract the features of folk music. Music is composed of many monosyllables. In psychology, sound includes the following four characteristics:

(1) pitch: pitch refers to people's feeling of the frequency of sound, determined by the number of vibration of an object;
(2) sound duration: sound duration refers to the duration of a note, which is determined by the duration of the vibration;
(3) sound intensity: sound intensity refers to the loudness that people feel, which is determined by the vibration amplitude;
(4) timbre: timbre refers to people's perception of sound quality, which is determined by the material, structure,. and shape of the sound body.

In the recognition of folk music, timbre is the main feature because music is played by different instruments. Timbre is a short-term feature, which can be extracted from the following three aspects.

### 2.2.1 Time-domain characteristics

Time-domain characteristics aim at the characteristics of the audio signal waveform. The time-domain features selected in this study are as follows.

(1) Short-time average energy (STE): it is used for reflecting the change of music signal amplitude. It refers to the average energy of the signal in the short-term audio window. For a short-time frame with a window length of $N$, suppose that the signal value of the $n$ -th sampling point is $x(n)$, the window function is represented by $w(n-m)$. For the $m$ -th frame, its STE can be expressed as:

$$E(m)=\frac{1}{N}\sum_{m}\big(x(n)w(n-m)\big)^2$$

(2) Zero crossing rate (ZCR): it refers to the number of times a signal waveform passes through the zero point in a frame. For the $m$ -th frame, its ZCR can be expressed as:

$$ZCR(m)=\frac{1}{2}\sum_{m}|sgn[x(n)]-sgn[x(n-1)]|w(n-m)$$

where $sgn$ stands for the sign function,

$$sgn=\begin{cases}1, x(n)\geq 0\\ 0, x(n)<0\end{cases}$$

### 2.2.2 Frequency-domain characteristics

Audio contains a lot of information, which needs to be obtained in the frequency domain analysis. The frequency domain features can be obtained by converting the signal to the frequency domain through Fourier transform. The features selected in this study are as follows.

(1) Spectrum centroid (SC): it refers to the characteristic quantity of the spectrum center of a signal. Fourier transform is represented by $F(\delta)$, $\delta\in(g,h)$, and the maximum and minimum values of frequency are represented by $g$ and $h$ respectively. Then SC can be expressed as:

$$SC=\frac{\sum_{\delta=g}^{h}\delta|F(\delta)|^2}{\sum_{\delta=g}^{h}|F(\delta)|^2}$$

(2) Spectrum energy (SE): it refers to the frequency domain energy of the signal, which can be expressed as:

$$SE=\sqrt{\frac{1}{h-g}\sum_{\delta=g}^{h}|F(\delta)|^2}$$

(3) Mel frequency cepstrum coefficient (MFCC) [6]: it refers to the cepstrum characteristics at Mel frequency, which has 13 dimensions. Suppose that the frequency of the music signal is $f$, then its Mel frequency is:

$$f_{mel}=2595\times log_{10}\left(1+\frac{f}{700}\right).$$

# 3 Support vector machine-based classification algorithm

SVM is a machine learning method [7], which has significant advantages in a small sample and nonlinear field and has been successfully applied in many fields, such as speech recognition [8] and image classification [9].

Suppose that in Euclidean space $R^d$, the training sample is $\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$ $(y\in\{+1,-1\})$, the linear discriminant function is $g(x)=wx+b$, and the classification plane equation is $wx+b=0$, where $w$ refers to the hyperplane normal vector, and $b$ refers to the offset. To separate the samples correctly, the problem can be expressed as:

$$min\frac{1}{2}\|w\|^2$$
$$y_i[(wx_i)+b]-1\geq 0$$

In the case of inseparable linearity, relaxation variable $\lambda$ and penalty factor $C$ are introduced. Then the above equation is transformed into:

$$min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\lambda_i$$
$$y_i[(wx_i)+b]\geq 1-\lambda_i$$

The Lagrange function is introduced to solve the above equation. Lagrange coefficient is set as $a_i$, then the optimal classification function is:

$$f(x)=sgn\left\{\sum_{i=1}^{N}a_iy_ik(x_i,x)+b\right\}$$

For any unclassified sample $x$, the result of classification can be obtained by calculating $f(x)$. $k(x_i, x_j)$ represents the kernel function. In SVM, the commonly used ones are:

(1) linear kernel function: $K(x_i, x_j) = x_i \cdot x_j$;

(2) polynomial kernel function: $K(x_i, x_j) = [(x_i \cdot x_j) + 1]^d$, where $d$ is an adjustable parameter;

(3) RBF kernel function: $K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{\sigma}\right)$, where $\sigma$ is an adjustable parameter.

In SVM, the RBF kernel function is the most commonly used and has the best performance; therefore, this study uses RBF kernel function. In SVM, the values of kernel function parameter $\sigma$ and penalty parameter $C$ have a great influence on the results [10], which needs to be determined in the experiment.

# 4 Experimental analysis

## 4.1 Folk music data set

The folk music was downloaded from the Internet and then converted to the WAV format of a single channel with a sampling frequency of 16 KHz by GoldWave software. The music file was processed by slicing by CoolEdit software and divided into 10 s segments. The final data sets obtained are shown in Table 2.

| Song | Types of folk music | Number |
|---|---|---|
| "Notturno in the Fisherboat", "Jackdaw Playing in the Water" | Zheng | 116 |
| "Ambush on All Sides", "Zhaojun Going Out of the Frontier" | Chinese lute | 121 |
| "Lofty Mountains and Flowing Water", "White Snow In Sunny Spring " | Guqin | 164 |
| "Journey to Suzhou", "Partridges Flying" | Bamboo flute | 138 |
| "Hundreds of Birds Worshipping the Phoenix", "A Flower " | Suona | 97 |
| "The Moon Over a Fountain", "The Song of Burying Flower" | Erhu | 167 |

Table 2: Data sets of folk music.

Features were extracted from the obtained data set, including 13-dimensional MFCC features and four one-dimensional features. The average value and standard deviation were taken, then each segment obtained 36-dimensional features. Then 80% of the features were selected as the training set, and 20% as the testing set.

## 4.2 Experimental results

Firstly, two parameters of SVM need to be determined. Two hundred of samples were selected. and determine the value of parameters through the cross test, as shown in Tables 3 and 4.

| $C$ | Recognition rate/% |
|---|---|
| $2^{-1}$ | 90.11 |
| 2 | 91.23 |
| $2^2$ | 93.87 |
| $2^3$ | 92.18 |
| $2^4$ | 92.09 |
| $2^5$ | 91.63 |
| $2^6$ | 91.29 |
| $2^7$ | 90.88 |
| $2^8$ | 90.64 |
| $2^9$ | 89.72 |
| $2^{10}$ | 88.33 |

Table 3: The influence of the value of $C$ on the recognition rate when the value of $\sigma^2$ takes 2.

| $\sigma^2$ | Recognition rate /% |
|---|---|
| $2^0$ | 88.64 |
| $2^1$ | 89.72 |
| $2^2$ | 90.07 |
| $2^3$ | 91.22 |
| $2^4$ | 92.08 |
| $2^5$ | 93.09 |
| $2^6$ | 93.87 |
| $2^7$ | 93.06 |
| $2^8$ | 92.18 |
| $2^9$ | 91.53 |
| $2^{10}$ | 90.27 |

Table 4: The influence of the value of $\sigma^2$ on the recognition rate when the value of $C$ takes 4.

It was seen from Tables 3 and 4 that the recognition rate of SVM was the highest when $C = 4$ and $\sigma^2 = 2^6$. Therefore, $C = 4$ and $\sigma^2 = 2^6$ were selected as the optimal parameters for the experiment.

The influence of feature selection on the results was compared. The selected features were time-domain, frequency-domain, and time + frequency-domain features of folk music. The results are shown in Figure 1.

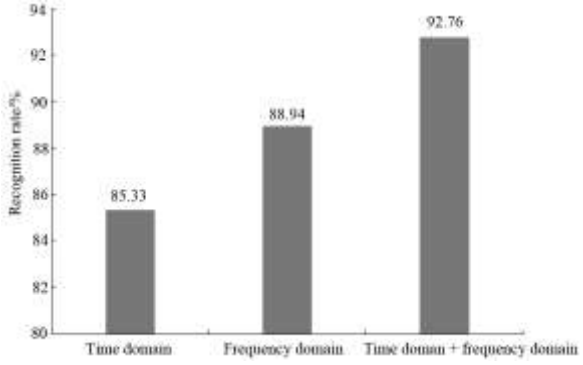



Figure 1: The influence of feature selection on the recognition rate.

It was seen from Figure 1 that the recognition rate of SVM was 85.33% when only the time domain features of folk music were selected and was 88.94% when only the frequency domain features were selected, and the increase of 4.23% might be due to the more feature dimensions contained in the frequency domain; when all the features were used for recognition, the recognition rate of SVM was 92.76%, which was 8.7% and 4.3% higher than the time domain and frequency domain. It was found that the recognition effect of SVM was good when all the features were used.

The recognition performance of SVM for different types of folk music is compared, and the results are shown in Figure 2.

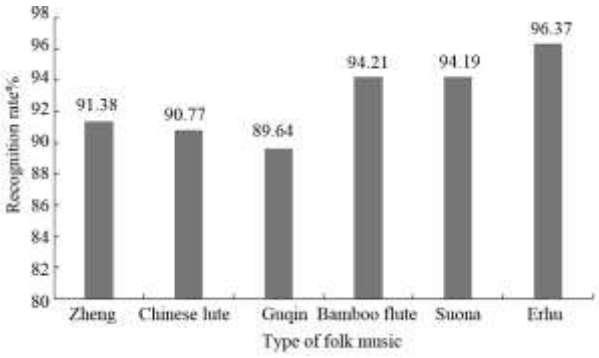



Figure 2: Recognition effect of different types of folk music.

It was seen from Figure 2 that SVM had the highest recognition rate for erhu, 96.37%, which might be because there was only one kind of string instrument, i.e., erhu, in the folk music data set studied in this study, which was significantly different from other types of folk music. The recognition rate of SVM was 91.38%, 90.77%, and 89.64% for Zheng, Chinese lute, and Guqin, which might be because the three instruments were slightly similar and more difficult to recognize.

To further verify the recognition performance of SVM, BP neural network (BPNN) [11], decision tree [12], and SVM were compared by the same folk music data set. The results are shown in Figure 2.

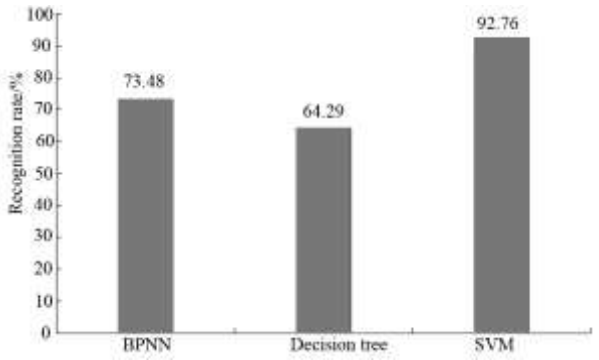



Figure 3: Comparison of recognition effects of different algorithms.

It was seen from Figure 3 that the recognition rates of the three algorithms were 73.48%, 64.29%, and 92.76%, respectively, and the recognition rate of SVM was 26.24% higher than that of BPNN and 44.28% higher than that of the decision tree. The results showed that SVM had significant advantages in the classification and recognition of folk music.

# 5 Discussion

The current research on music recognition and classification includes the classification of genres [13], musical instruments [14], emotions [15], composers, and so on. Through the identification and classification, users can quickly and accurately retrieve the music they want to hear, and it is also more convenient to manage the music. With the development of technology, music recognition and classification has made great progress, and more and more machine learning methods have been applied, such as hidden Markov, decision tree, nearest neighbor, etc. [16]. In this study, SVM was used for classifying folk music.

In the identification and classification of folk music, this study extracted the time-domain and frequency-domain features to form the folk music data set and then used the SVM method for classification. In the experiment, to obtain the optimal parameters of SVM, this study analyzed the influence of different values on the results by the cross-check method, and then the obtained optimal parameters were used for the next step of the experiment. The results showed that the recognition rate of SVM was higher when more comprehensive features were selected. In folk music recognition, when using time-domain and frequency-domain features, the recognition rate of SVM reached 92.76%. In recognizing different types of folk music, the recognition rate of SVM for erhu was the highest (96.37%), while the recognition rates of three plucked instruments were relatively low. In comparison with other methods, this study selected BPNN and decision tree for comparison. It was seen from Figure 2 that the recognition rate of SVM used in this study was significantly higher than the other two methods, which indicated that SVM had a better performance in the recognition of folk music.

Although some achievements have been made in this paper, further research is needed. In future work, we will:

(1) further study the selection of features;
(2) further improve the classification performance of SVM;
(3) perform experiments on a more extensive data set.

# 6 Conclusion

In this study, the method of feature extraction was analyzed for the recognition and classification of folk music, SVM was selected as the classifier, and a data set was established for experimental analysis. The results demonstrated that:

(1) the selection of parameters had an influence on the result of folk music recognition;
(2) when all the features were used, the recognition rate of SVM was the highest (92.76%);

(3) SVM had the highest recognition rate for erhu, reaching 96.37%;
(4) compared with BPNN and decision tree, SVM had a significantly higher recognition rate.

# References


[1] Huang W, Zhang YT. (2020). Application of Hidden Markov Chain and Artificial Neural Networks in Music Recognition and Classification. *ICCDE 2020: 2020 The 6th International Conference on Computing and Data Engineering*, pp. 49-53.

[2] Abidin D, Özacar T, Ozturk O. (2018). Using classification algorithms for Turkish music makam recognition. 6, pp. 377-393. https://doi.org/10.15317/Scitech.2018.139

[3] Rao Z, Feng C. (2018). Sparse representation classification-based automatic chord recognition for noisy music. *Journal of Information Hiding and Multimedia Signal Processing*, 9, pp. 3400-409.

[4] Iloga S, Romain O, Tchuente M. (2018). A sequential pattern mining approach to design taxonomies for hierarchical music genre recognition. *Pattern Analysis & Applications*, 21, pp. 3363-380.

[5] Xie CY. (2015). *Research on the Development of National Music in the New Media Era*, International Conference on Education. Atlantis Press.

[6] Lalitha S, Geyasruti D, Narayanan R, Shravani M. (2015). Emotion Detection Using MFCC and Cepstrum Features. *Procedia Computer Science*, 70, pp. 329-35. https://doi.org/10.1016/j.procs.2015.10.020

[7] Abdiansah A, Wardoyo R. (2015). Time Complexity Analysis of Support Vector Machines (SVM) in LibSVM. *International Journal of Computer Applications*, 128, pp. 975-8887. https://doi.org/10.5120/ijca2015906480

[8] Bhavan A, Chauhan P, Hitkul, Shah RR. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge Based Systems*, 184, pp. 104886. https://doi.org/10.1016/j.knosys.2019.104886

[9] Gao L, Li J, Khodadadzadeh M, Plaza A. (2015). Subspace-Based Support Vector Machines for Hyperspectral Image Classification. *IEEE Geoscience & Remote Sensing Letters*, 12, pp. 349-353. https://doi.org/10.1109/LGRS.2014.2341044

[10] Rosales-Pérez A, Gonzalez J A, Coello CAC, Escalante HJ, Reyes-Garcia CA. (2015). Surrogate-assisted multi-objective model selection for support vector machines. *Neurocomputing*, 150, pp. 163-172. https://doi.org/10.1016/j.neucom.2014.08.075

[11] Wang J, Yan WQ. (2016). BP-Neural Network for Plate Number Recognition. *International Journal of Digital Crime & Forensics*, 8, pp. 34-45.

[12] Kumar R, Singh B, Shahani DT, Chandra A. (2015). Recognition of Power Quality events using S-transform based ANN classifier and rule based decision tree. *IEEE Transactions on Industry Applications*, 51, pp. 1249-1258.

[13] Costa YMG, Oliveira LS, Silla CN. (2017). An Evaluation of Convolutional Neural Networks for Music Classification Using Spectrograms. *Applied Soft Computing*, 52, pp. 28-38. https://doi.org/10.1016/j.asoc.2016.12.024

[14] Giannoulis D, Klapuri A. (2013). Musical Instrument Recognition in Polyphonic Audio Using Missing Feature Approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 21, pp. 1805-1817. https://doi.org/10.1109/TASL.2013.2248720

[15] Bai J, Luo K, Peng J, Shi J, Wu Y, Feng L, Li J, Wang Y. (2017). Music Emotions Recognition by Machine Learning With Cognitive Classification Methodologies. *International Journal of Cognitive Informatics and Natural Intelligence*, 11, pp. 80-92. https://doi.org/10.4018/IJCINI.2017100105

[16] Nasridinov A, Park YH. (2014). A Study on Music Genre Recognition and Classification Techniques. *International Journal of Multimedia & Ubiquitous Engineering*, 9, pp. 31-42. https://doi.org/10.14257/ijmue.2014.9.4.04

# CONTENTS OF *Informatica* Volume 44 (2020) pp. 1–529

## Papers

AMPOMAH, E.K. & , Z. QIN, G. NYAME, F.E. BOTCHEY. 2020. Stock Market Decision Support Modeling with Tree-based AdaBoost Ensemble Machine Learning Models. Informatica 44:477–489.

BALI, M. & , A. TARI, A. ALMUTAWAKEL, O. KAZAR. 2020. Smart Design for Resources Allocation in IoT Application Service Based on Multi-agent System and CSP. Informatica 44:373–386.

BARREIROS, A. & , J.B. CARDOSO. 2020. Design Optimization Average-Based Algorithm. Informatica 44:23–33.

BATHLA, Y. & , S. SZENASI. 2020. A Web Server to Store the Modeled Behavior Data and Zone Information of the Multidisciplinary Product Model in the CAD Systems. Informatica 44:275–283.

BOZINOVSKI, S. & . 2020. Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. Informatica 44:291–302.

CARLSEN, L. & . 2020. The Iris Dataset Revisited - a Partial Ordering Study. Informatica 44:35–44.

DALILA, C. & , E.A.O. BADIS, B. SADDEK, N.-A. AMINE. 2020. Feature Level Fusion of Face and Voice Biometrics Systems Using Artificial Neural Network for Personal Recognition. Informatica 44:85–96.

DEY, A. & . 2020. Probabilistic Weighted induced Multi-Class Support Vector Machines for Face Recognition. Informatica 44:459–467.

DINH, H.M. & , D.V. NGUYEN, L.V. TRUONG, T.P. DO, T.T. PHAN, N.D. NGUYEN. 2020. Cycle Time Enhancement by Simulated Annealing for a Practical Assembly Line Balancing Problem. Informatica 44:127–138.

DJEBBAR, A.M. & . 2020. A Hybrid Discrete Artificial Bee Colony for the Green Pickup and Delivery Problem with Time Windows. Informatica 44:507–519.

GAMS, M. & . 2020. Call for Special Issue of Electronics. Informatica 44:287–288.

GARCÍA-ZURDO, R. & . 2020. Creation of Facial Composites from User Selections using Image Gradient. Informatica 44:15–22.

GRIGORAS, N. & . 2020. Minimum Flows in Parametric Dynamic Networks the Static Approach. Informatica 44:303–310.

GROZNIK, V. & . 2020. Artificial Intelligence Methods for Modelling Tremor Mechanisms. Informatica 44:285–286.

GRŮBER, M. & , J. MATOUŠEK, Z. HANZLÍČEK, D. TIHELKA. 2020. Dialogue Act-Based Expressive Speech Synthesis in Limited Domain for the Czech Language. Informatica 44:147–165.

HARIATI, M. & . 2020. Formal Verification Issues for Component-Based Development. Informatica 44:469–475.

JENA, M. & , S. DEHURI. 2020. DecisionTree for Classification and Regression: A State-of-the Art Review. Informatica 44:405–420.

KABASSI, K. & , A. BOTONIS, C. KARYDIS. 2020. Evaluating Websites of Conservation Labs in Museums using Fuzzy Multi-Criteria Decision Making Theories. Informatica 44:45–54.

KARNA, H. & , S. GOTOVAC, L. VICKOVIĆ. 2020. Data Mining Approach to Effort Modeling On Agile Software Projects. Informatica 44:231–239.

KAZAKOVTSEV, L.A. & . 2020. Application of Algorithms with Variable Greedy Heuristics for k-Medoids Problems. Informatica 44:55–61.

KOHEK, Š. & . 2020. Interactive Synthesis and Visualisation of Vast Areas with Geometrically Diverse Trees. Informatica 44:109–110.

LUCKY, L. & , A.S. GIRSANG. 2020. Hybrid Nearest Neighbors Ant Colony Optimization for Clustering Social Media Comments. Informatica 44:63–74.

MILIĆ, D.C. & , Z. KRPIC, F. SUŠAC. 2020. E-learning in Business Practice, a Case Study During COVID-19 in Croatia. Informatica 44:427–436.

MISHRA, S. & , S.K. MISHRA. 2020. Performance Assessment of a Set of Multi-Objective Optimization Algorithms for Solution of Economic Emission Dispatch Problem. Informatica 44:349–360.

NEMMICH, M.A. & , F. DEBBAT, M. SLIMANE. 2020. Hybrid Bees Approach based on Improved Search Sites Selection by Firefly Algorithm for Solving Complex Continuous Functions. Informatica 44:183–198.

NGUYEN, S.C. & , K.H. HA, H.M. NGUYEN. 2020. A Robust Image Watermarking Scheme Based on the Laplacian Pyramid Transform. Informatica 44:75–84.

NI, H. & . 2020. Face Recognition Based on Deep Learning Under the Background of Big Data. Informatica 44:491–495.

PANDA, D. & , P.S.R. DASH, R. RAY, S. PARIDA. 2020.

Predicting the Causal Effect Relationship Between COPD and Cardio Vascular Diseases. Informatica 44:447–457.

PANDA, M. & , S. DEHURI, A.K. JAGADEV. 2020. Multi-Objective Artificial Bee Colony Algorithms and Chaotic-TOPSIS Method for Solving Flowshop Scheduling Problem and Decision Making. Informatica 44:241–262.

PENG, R. & , Y. YAO. 2020. Comparison of Community Structure Partition Optimization of Complex Networks by Different Community Discovery Algorithms. Informatica 44:97–102.

PHAM, V.-A. & , D.-H. HOANG, H.-H. CHUNG-NGUYEN, M.-K. TRAN, M.-T. TRAN. 2020. Privacy Preserving Visual Log Service with Temporal Interval Query using Interval Tree-based Searchable Symmetric Encryption. Informatica 44:115–125.

PISANSKI, T. & , M. PISANSKI, J. PISANSKI. 2020. A Novel Method for Determining Research Groups from Co-authorship Network and Scientific Fields of Authors. Informatica 44:139–146.

RAMOU, N. & , N. CHETIH, Y. BOUTICHE, R. ABDELKADER. 2020. Automatic Image Segmentation for Material Microstructure Characterization by Optical Microscopy. Informatica 44:367–372.

ROCHE, M. & . 2020. How to Define Co-occurrence in a Multidisciplinary Context?. Informatica 44:387–393.

ROY, A. & . 2020. Designing Hybrid Intelligence Based Recommendation Algorithms: An Experience Through Machine Learning Metaphor. Informatica 44:401–402.

SAHU, M. & , D.P. MOHAPATRA. 2020. Computing Dynamic Slices of Feature-Oriented Programs with Aspect-Oriented Extensions. Informatica 44:199–224.

SAIFAN, R. & , K. SHARIF, M. ABU-GHAZALEH, M. ABDELMAJEED. 2020. Investigating Algorithmic Stock Market Trading Using Ensemble Machine Learning Methods. Informatica 44:311–325.

SALIMZADEH, S. & , S. KANDULU. 2020. Teeth Segmentation of Bitewing X-Ray Images Using Wavelet Transform. Informatica 44:421–426.

SHIJINA, V. & , U. ADITHYA, J.J. SUNIL. 2020. Similarity Measure of Multiple Sets and its Application to Pattern Recognition. Informatica 44:335–348.

SIMONAK, S. & . 2020. Increasing the Engagement Level in Algorithms and Data Structures Course by Driving Algorithm Visualizations. Informatica 44:327–334.

SU, Y. & . 2020. Research on Recognition Algorithm of Important Nodes in Complex Network. Informatica 44:103–107.

UTKIN, L.V. & . 2020. Improvement of the Deep Forest Classifier by a Set of Neural Networks. Informatica 44:1–13.

VÁRKONYI, G.G.D. & , A. GRADISEK. 2020. Data Protection Impact Assessment Case Study for a Research Project Using Artificial Intelligence on Patient Data. Informatica 44:497–505.

VERA, J.C.D. & , G.M.N. ORTIZ, C. MOLINA, M.A. VILA. 2020. Knowledge Redundancy Approach to Reduce Size in Association Rules. Informatica 44:167–182.

WANG, D. & , G. XU. 2020. Research on the Detection of Network Intrusion Prevention With Svm Based Optimization Algorithm. Informatica 44:269–274.

WANG, H. & . 2020. Research on Data Transmission Optimization Of Communication Network Based on Reliability Analysis. Informatica 44:361–365.

WANG, X. & . 2020. Research on Recognition and Classification of Folk Music Based on Feature Extraction Algorithm. Informatica 44:521–525.

WOODS, N.C. & , C.A. ROBERT. 2020. Colour-Range Histogram technique for Automatic Image Source Detection. Informatica 44:225–230.

XU, S. & . 2020. Association Rule Model of On-demand Lending Recommendation for University Library. Informatica 44:395–399.

YUSIONG, J.P.T. & , P.C. NAVAL. 2020. A Semi-Supervised Approach to Monocular Depth Estimation, Depth Refinement, and Semantic Segmentation of Driving Scenes using a Siamese Triple Decoder Architecture. Informatica 44:437–445.

ZHANG, R. & , W. SHI. 2020. Research on Resource Allocation and Management of Mobile Edge Computing Network. Informatica 44:263–268.

## Editorials

BINH, H.T.T. & , I. IDE. 2020. Introduction to Special Issue "SoICT 2019". Informatica 44:113–113.

# 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 scientific 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 temperature, 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, energy 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 departments 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 universities 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; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean 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 strategy for technological development to foster synergies between research and industry, to promote joint ventures between 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 philosophical 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 published within one year of receipt of email with the text in Informatica 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 editor.

# 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 twentysix 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 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 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, educational and governmental institutions. Others should subscribe.

## Informatica WWW:

## http://www.informatica.si/

**Referees from 2008 on:**

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Čosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

# *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 2020 (Volume 44) 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: Borut Žnidar, borut.znidar@gmail.com.
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 44 Number 4 December 2020** **ISSN 0350-5596**

# *Informatica*

## An International Journal of Computing and Informatics


| | | |
|---|---|---|
| DecisionTree for Classification and Regression: A State-of-the Art Review | M. Jena, S. Dehuri | **405** |
| Teeth Segmentation of Bitewing X-Ray Images Using Wavelet Transform | S. Salimzadeh, S. Kandulu | **421** |
| E-learning in Business Practice, a Case Study During COVID-19 in Croatia | D.C. Milić, Z. Krpić, F. Sušac | **427** |
| A Semi-Supervised Approach to Monocular Depth Estimation, Depth Refinement, and Semantic Segmentation of Driving Scenes using a Siamese Triple Decoder Architecture | J.P.T. Yusiong, P.C. Naval | **437** |
| Predicting the Causal Effect Relationship Between COPD and Cardio Vascular Diseases | D. Panda, P.S.R. Dash, R. Ray, S. Parida | **447** |
| Probabilistic Weighted induced Multi-Class Support Vector Machines for Face Recognition | A. Dey | **459** |
| Formal Verification Issues for Component-Based Development | M. Hariati | **469** |
| Stock Market Decision Support Modeling with Tree-based AdaBoost Ensemble Machine Learning Models | E.K. Ampomah, Z. Qin, G. Nyame, F.E. Botchey | **477** |
| Face Recognition Based on Deep Learning Under the Background of Big Data | H. Ni | **491** |
| Data Protection Impact Assessment Case Study for a Research Project Using Artificial Intelligence on Patient Data | G.G.D. Várkonyi, A. Gradišek | **497** |
| A Hybrid Discrete Artificial Bee Colony for the Green Pickup and Delivery Problem with Time Windows | A.M. Djebbar | **507** |
| Research on Recognition and Classification of Folk Music Based on Feature Extraction Algorithm | X. Wang | **521** |

Informatica **44** (2020) Number 4, pp. 405–529