SUPERVISED LEARNING - A SYSTEMATIC LITERATURE REVIEW

A PREPRINT

Salim Dridi
contact.salimdridi@gmail.com - salimdridi.info

December 28, 2021

ABSTRACT

Machine Learning (ML) is a rapidly emerging field that enables a plethora of innovative approaches to solving real-world problems. It enables machines to learn without human intervention from data and is used in a variety of applications, from fraud detection to recommendation systems and medical imaging. Supervised learning, unsupervised learning, and reinforcement learning are the 3 main categories of ML. Supervised learning involves pre-training the model on a labeled dataset and entails two distinct types of learning: classification and regression. Regression is used when the output is continuous. By contrast, classification is used when the output is categorical.

Supervised learning aims to optimize class label models using predictor features. Following that, a second classifier is used to assign class labels to the test data in cases where the values of the predictor characteristics are known but the value of the class label is unknown. In classification, the label identifies the class to which the training set belongs. However, in regression, the label is a real-value response that corresponds to the example.

Numerous supervised learning approaches and algorithms have been proposed: XGBoost, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbor to name a few. This survey paper examines supervised learning by offering a thorough assessment of approaches and algorithms, performance metrics, and the merits and demerits of numerous studies. This paper will point researchers in new directions and enable them to compare the efficacy and effectiveness of supervised learning algorithms.

Keywords Supervised Learning · Literature Review · Survey · ML · Supervised Learning approaches · Classification and Regression

Table of Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

ML, due to the concepts it inherits, can be regarded a subfield of Artificial Intelligence (AI). It enables prediction; for this purpose, its basic building blocks are algorithms. ML enables systems to learn on their own rather than being explicitly programmed to do so, resulting in more intelligent behavior. It generates data-driven predictions by developing models that discover patterns in historical data and utilize those patterns to generate predictions [1] [2]. The general architecture of ML consists of several steps: business understanding (understanding and knowledge of the domain), data acquisition and understanding (gathering and understanding data), modeling (which entails feature engineering, model training, and evaluation), and deployment (deploy the model on the cloud).

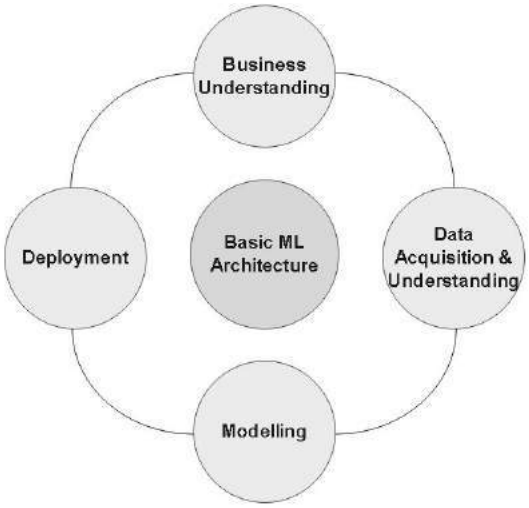

Figure 1: Basic Architecture of ML

Supervised Learning, Unsupervised Learning, and Reinforcement Learning are the three major categories of ML. In Supervised Learning, the model is trained on labeled data and is then used to generate predictions on unlabeled data [3]. In unsupervised learning, a model is trained on unlabeled data and said model automatically learns from that data by extracting features and patterns from it. In Reinforcement Learning, an agent is trained on the environment and this enables said agent to find the optimum solution and accomplish a goal in a complex situation [4].

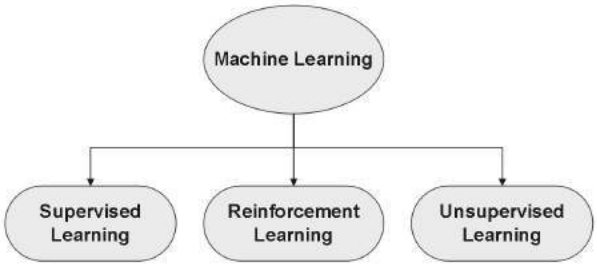

Figure 2: Major Categories of ML

1.1 Supervised Learning

ML algorithms treat each instance of a dataset as a collection of features. These features may be binary, categorical, or continuous in nature. If the instances are labeled, then this type of learning is termed as supervised learning [2]. Supervised Learning involves training the model on labeled data and testing it on unlabeled data. Its fundamental architecture begins with dataset collection; the dataset is then partitioned into testing and training data; and then, the data is preprocessed. Extracted features are fed into an algorithm and the model is then trained to learn the features

associated with each label. Finally, the model is supplied with the test data, and said model makes predictions on the test data by providing the expected labels, as illustrated in figure 3.

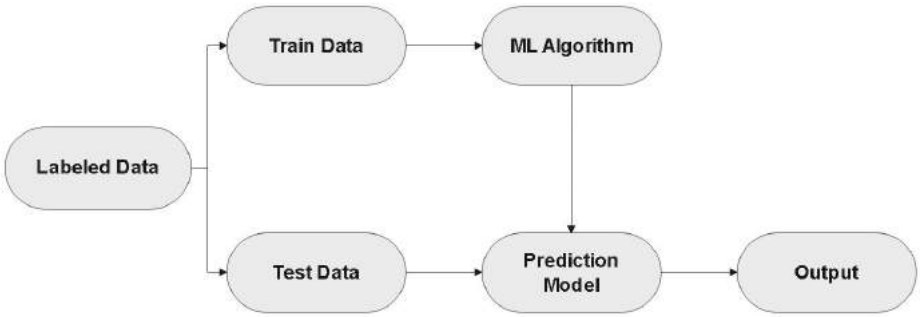

Figure 3: Basic Architecture of Supervised Learning

Classification and regression are the two broad types of supervised learning. Both are discussed in detail in the following sections.

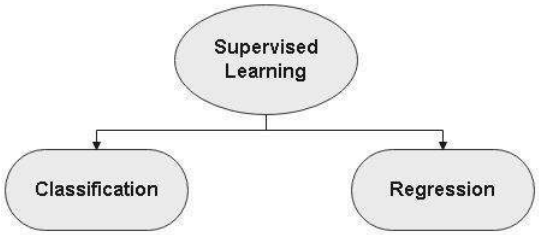

Figure 4: Types of Supervised Learning

1.1.1 Classification

In classification, a model predicts unknown values (set of Outputs) based on a set of known values (set of inputs) [2]. When the output is in categorical form, the problem is referred to as a classification one [5]. Generally, in classification, a dataset's instances are categorised according to specified classes [5] [6]. Classification can be applied to both structured and unstructured datasets. Some terms used in Classification are: Classification model, Classification algorithm, and feature. A classification algorithm, alternatively referred to as a classifier, learns from the training dataset and assigns each new data point to a certain class. In comparison, a classification model uses a mapping function, which is concluded by said model from the training dataset, to predict the class label for the test data. Finally, a feature is associated with the dataset, which helps in building a precise predictive model.

The classification process is depicted in Figure 5. Data collection and preprocessing are the first steps in building a classification model. Preprocessing is the process of cleansing data by eliminating noise and duplicates. Numerous techniques are used to preprocess data, among which "brute-force" is the simplest and most common one. The data is then split into train and test sets using the cross-validation technique. The next stage is to train the model using class labels; in Python, the sci-kit-learn package has a function called "fit-transform (X,Y)" that maps X (input data) to Y (labels) for the purpose of preparing the classifier. The next step is to forecast the new dataset's class or label. Finally, the classification algorithm is evaluated using the test data.

There are two distinct classification types: binary and multi-label [5]. Binary Classification is used when the outcome is binary or has two classes. For instance, in an ambiguity detection process, the model predicts whether or not sentences are ambiguous and as a result, there are only two possible outcomes/classes; this is referred to as binary classification. However, multi-label classification is made up of a distribution of classes. For example, in predicting mental disorders, there are multiple ones such as depression, anxiety, schizophrenia, bipolar disorder, and

Post-traumatic Stress Disorder (PTSD). Thus, the outcome can fall into one of these five categories; this is termed multi-label classification.

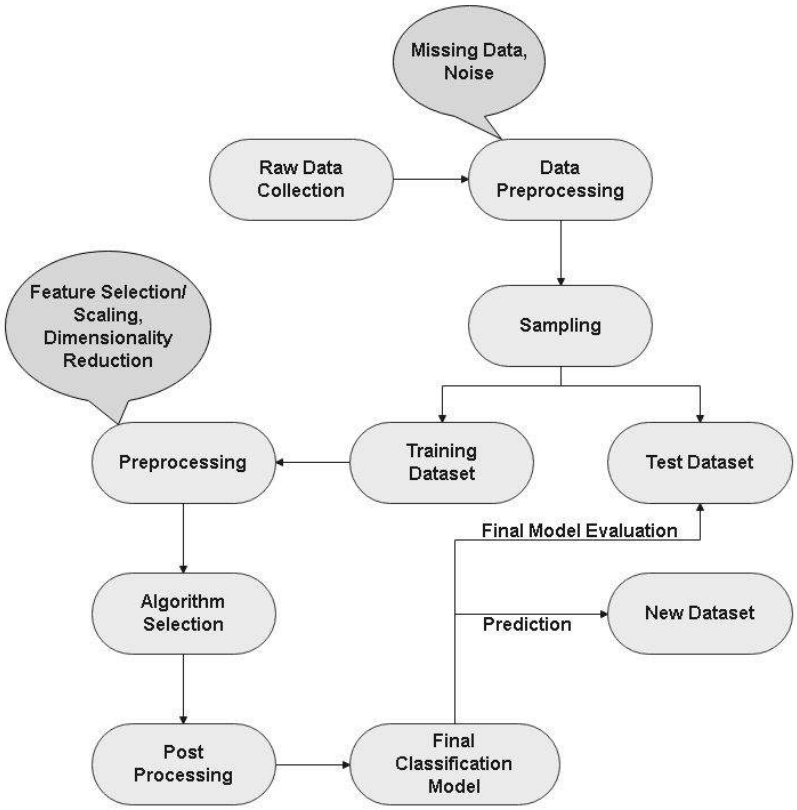

Figure 5: Workflow of Classification in Supervised Learning

1.1.2 Regression

Regression is a supervised learning technique that permits the discovery of correlations between variables and the prediction of continuous values based on these variables. When the output is continuous, the problem is referred to as a regression one [5]—for instance, predicting a person's weight, age, or salary, weather forecasting, or housing price forecasting. In regression, X (input variables) is mapped to Y (continuous output). Classification is the process of predicting the discrete labels of the input. Regression, on the other hand, is concerned with the prediction of continuous values. Regression is divided into two main categories: Simple Linear and Multiple (figure 6). In simple linear regression, a straight line is drawn to define the relationship between two variables (X and Y). In contrast, Multiple regression encompasses multiple variables and is further divided into linear and non-linear.

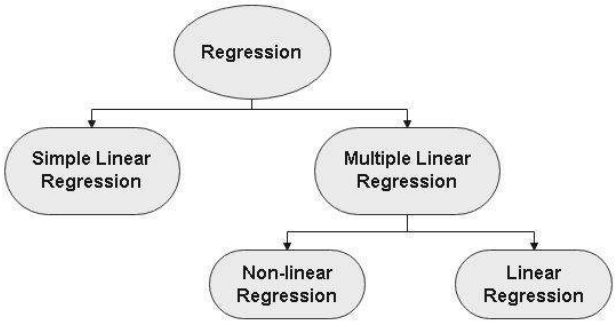

Figure 6: Types of Regression

The primary goal of this literature review is to discover and assess works on the stated subject. This review will assist researchers in identifying future research areas by providing an overview of Supervised Learning Approaches and Algorithms, the metrics used to evaluate the performance of each supervised learning model, and a comparative analysis of the accuracy of each supervised learning model.

The following section presents the literature review of the topic, Section 3 provides the proposed methodology for conducting this systematic literature review, Section 4 provides the results of the study and the related discussion, and Section 5 presents the conclusion.

2 Literature Review

During the previous decade, numerous researchers have already performed surveys on supervised learning. For instance, the authors of paper [7] wrote a survey on various supervised learning classification approaches. Their study examined five classification methods: Naïve Bayes, Neural Network, Decision tree, Support vector machine, and K-Nearest neighbor; and presented a taxonomy of each paper's benefits and shortcomings. Additionally, they categorised the papers according to their research topic, classification algorithm, and publication year. Their survey includes articles in a variety of disciplines, including medicine, agriculture, education, business, and networking. According to their research, the most frequently used classification algorithms are decision trees and Naïve Bayes. Their survey, however, was limited to only five classification strategies.

The authors of study [2] classified supervised learning techniques into five categories: Logic-based algorithms, Statistical learning algorithms, Instance-based learning, Support vector machines, and deep learning. Additionally, they demonstrated the general pseudocodes of decision trees, rule learners, Bayesian networks, and instance-based learners. According to their survey, neural networks and SVM outperform other algorithms when dealing with continuous data. In comparison, logic-based algorithms perform better when the data is categorical. Additionally, they stated that Naïve Bayes is capable of doing well with small datasets. On the other hand, SVM and neural networks require large datasets to reach optimal accuracy. However, their research focused exclusively on classification algorithms and didn't cover regression.

Similarly, [5] provides an overview of supervised classification approaches and classifies them as: logically learning algorithms, support vector machine, statistically based algorithms, and lazy learning algorithms. This survey defines, details, and discusses the benefits, drawbacks, and applications of each technique. The authors also conducted a comparative analysis of the accuracy of four widely used algorithms: SVM, Naïve Bayes, Decision tree, and k-NN; at the conclusion of the paper using a dataset from the Census Bureau Database. Several comparison parameters such as classification speed, learning speed, and noise tolerance were used in their analysis. According to their findings, SVM at 84.94% outperformed k-NN, Naïve Bayes, and Decision Trees in terms of accuracy, respectively.

In paper [7], the authors reviewed supervised text classification techniques. Their survey covered three machine learning approaches: NB, SVM, and k-NN; and the performance evaluation measures associated with each. Additionally, they mentioned several weighting methods for text classification. As a result, k-NN outperformed the other ML algorithms. According to this study, the performance of the algorithms is dependent on the dataset, i.e. each algorithm performed differently on different datasets. Unfortunately, this study was limited to three classification models.

Study [8] compares supervised learning methods empirically. The authors use eight comparison parameters to contrast the following supervised learning algorithms: SVM, ANN, Logistic regression, Naive Bayes, k-NN, Decision tree, Random Forest, Bagged trees, memory-based learning, and Boosted stumps. Those parameters were: Accuracy, precision and recall, F-score, Cross entropy, ROC Curve, Squared error, average precision, and breakeven point. According to their findings, calibrated boosted trees outperformed all by scoring highly on all comparison parameters. Random forest came in a close second place, followed by SVM. The performance of logistic regression, Naïve Bayes, and decision trees was, however, poor. It is worth noting that the models' calibration is surprisingly effective at producing an excellent performance.

While researchers have made significant attempts at conducting surveys in the domain of supervised learning, there is still a need for a systematic literature review (SLR), which is one of the main issues of the aforementioned studies. Additionally, these surveys are restricted to a subset of classification methods, further justifying an enhanced and organized SLR.

3 Methodology of this Systematic Literature Review

A SLR is a type of study that tries to identify and analyze existing literature on a certain subject. It is conducted formally and methodically, and is objective and reproducible. We conducted this SLR using the principles established by Barbara Ann Kitchenham [9] for performing SLRs.

This SLR will follow Kitchenham's guidelines and well-defined steps. The first stage is to define the research questions and then analyze the collected data to answer them. The second phase entails defining the search procedure. This study has included conference proceedings and journal articles dating back to 2011. "IEEE Xplore," "the Association for Computing Machinery (ACM)," and "Science Direct Elsevier" were the three databases used to look for papers. In the third phase, we defined the search strings that were used to retrieve related works from these databases. Then, the papers were subjected to inclusion and exclusion criteria: we included papers that were relevant to the topic and have a publication date no older than 2011 and excluded all others. Finally comes Data Extraction which involves extracting information from the selected list of articles that address the three research questions.

3.1 Research Questions

The following are the Research Questions (RQs) formulated for this study.

RQ1: What are the approaches/algorithms that are used for problem-solving in supervised learning?

RQ2: What are the widely used evaluation metrics for measuring the performance of the employed supervised learning model?

RQ3: What are the merits and demerits of each study in supervised learning?

3.2 Search Strategy

Once the key terms were identified, we initiated the search process. From these different terms, we formulated the query strings.

3.3 Query Strings

To formulate query strings, we followed the three databases' (ACM Digital Library, IEEE Xplore, and Science Direct Elsevier) guidelines in using Boolean operators, synonyms, and different terms. For our initial search strings, thousands of papers appeared, so we had to refine them and use the Advanced Search option so a lower number of papers would be returned to us. Table 1 below shows the final Search strings that were used.

Database Name	Search String
IEEE Xplore	((supervised learning) AND (Supervised ML approaches))
ACM Digital Library	acmdlTitle:(supervised learning) AND (+"algorithms and approaches")
Science Direct	Supervised learning algorithms AND Supervised learning approaches

Table 1: Search Strings

3.4 Search Results

To retrieve the articles, we ran the queries in August 2021. The result of the query strings is shown in Table 2. Then, for the citations and bibliographies, we imported these papers into Mendeley Library. After eliminating duplicates, we ended up with a total of 139 articles.

Database Name	Search String	Years	Number of Papers
IEEE Xplore	((supervised learning) AND (Supervised ML approaches))	2011-2021	51
ACM Digital Library	acmdlTitle:(supervised learning) AND (+"algorithms and approaches")	2011-2021	41
Science Direct	Supervised learning algorithms AND Supervised learning approaches	2011-2021	55

Table 2: Search Results

3.5 Study Selection Criteria

Based on our stated research questions and SLR topic, inclusion and exclusion criteria were defined as follows.

Inclusion Criteria

Papers were included based on the following assumptions:

1. Those which were published between 1st January 2011 and 1st August 2021.
2. Those which contain a supervised learning approach.
3. Those that discuss classification and regression algorithms in the domain of supervised learning.

Exclusion Criteria

Papers were excluded based on the following assumptions:

1. Those which were published before 2011.
2. Those which do not focus on supervised ML.
3. Those which are describing supervised learning but did not focus on any approach or algorithm of supervised learning.

3.6 Search Process

We used the following keywords to conduct our search in the aforementioned databases: "supervised machine learning," "supervised learning algorithm," and "supervised learning approach." Then, we set the year filter to exclude works published prior to 2011. Following that, we assessed the papers that were relevant to our investigation. Several were eliminated solely on the Title, while others were eliminated after reading the Abstract. Finally, as shown in Table 3, we ended up with 57 papers for our survey.

Database Name	Years	No. of Papers	No. of Papers Excluded on Basis of Title	No. of Papers Excluded on Basis of Abstract	No. of Papers Selected
IEEE Xplore	2011-2021	51	14	18	19
ACM Digital Library	2011-2021	41	12	13	16
Science Direct	2011-2021	55	16	17	22

Table 3: Number of Papers Selected After Applying Inclusion and Exclusion Criteria

Figure 7 summarizes the strategy used to conduct this literature review on supervised learning. As noted previously, three databases were used to choose the publications. Following that, we obtained a total of 139 articles from three databases. Then, as discussed, we applied the inclusion and exclusion criteria. After applying the inclusion and exclusion criteria, 57 papers remained that also matched our ten-year time window (1 January 2011 - 1 August 2021). The next section contains the results of the study.

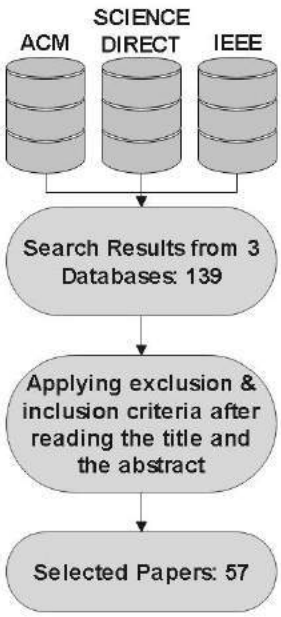

Figure 7: Methodology Used for this Systematic Literature Review on Supervised ML

4 Results and Discussion

In this section, the results corresponding to each formulated research question are provided.

RQ1: What are the approaches/algorithms that are used for problem-solving in supervised learning?

Many supervised learning approaches and algorithms have been proposed since the last decade. Our survey divides the supervised learning them into five categories: logic-based, statistics-based, instance-based, support vector machines, and deep learning. Figure 8 lays out the different approaches in supervised learning, whereas Figure 9 presents an overview of the algorithms.

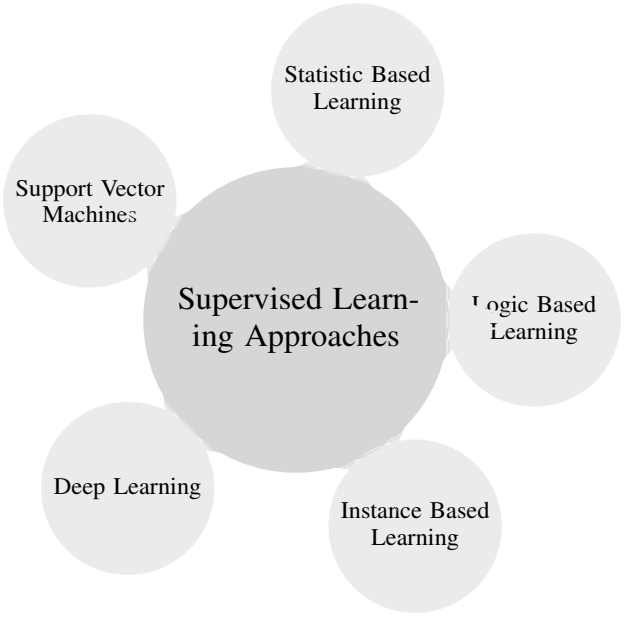

Figure 8: Supervised Learning Approaches

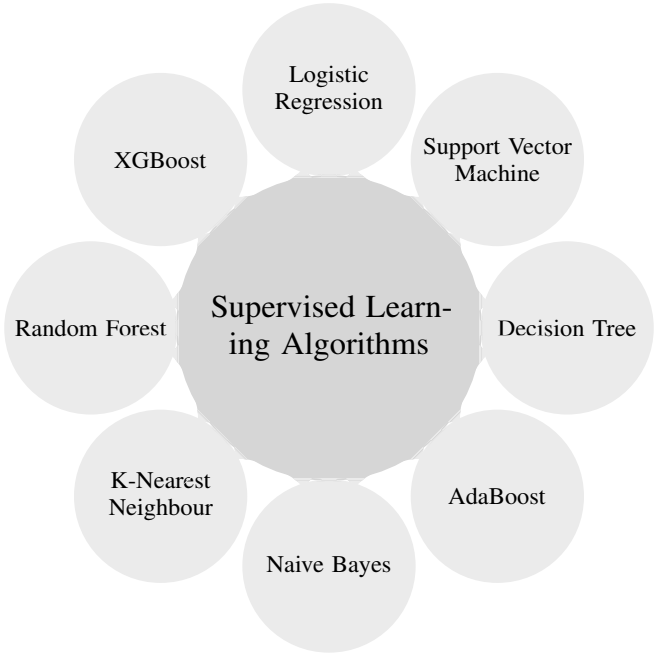

Figure 9: Widely Used Supervised Learning Algorithms

4.1 Statistics Based Learning

The statistics-based approach simplifies the problem through the use of distributive statistics. The prediction task is based on the structure of the distribution. The statistical-based approach to learning involves the Naïve Bayes Algorithm.

4.1.1 Naïve Bayes (NB)

Naïve Bayes (NB) is a popular algorithm for classification and regression in predictive modeling. NB is based on the Bayes theorem, with features assumed to be independent. NN is an acyclic graph with a single parent and many children,

all of whom are independent. Numerous studies used NB for a variety of classification tasks. In [10], the authors proposed an opinion-based book recommendation system by employing NB classification to classify and summarize customer feedback on a book. NB performed well in recommending top-ranked books to customers. The authors of [11] utilized NB to classify whether email was spam or legitimate. The results showed that NB performed better than other algorithms when it categorized about 95% of users' spam emails correctly. In Study [12], deep feature weighting NB was used to classify Chinese text. According to the findings of the authors, deep feature NB outperformed simple NB. In [13], the authors presented a system for emotion recognition based on audio signals. Aspects of audio signals such as pitch, ZCR, and energy were classified using NB. The authors of [14] proposed a method for classifying semantic web services using NB. Finally, in [15], the authors described a patient-centric clinical decision support system based on the NB classifier that maintained patient anonymity. The concept offered increased diagnostic precision and minimized diagnostic time.

The pseudocode of the Naïve Bayes algorithm [56] is:

Algorithm 1 Pseudo-Code of Naïve Bayes

Inputs:
Training dataset T,
F=(f1,f2,f3,..,fn) //value of the predictor variable in testing dataset
Output:
A class of testing dataset

1: Read the training dataset T
2: Calculate the mean and standard deviation of the predictor variable in each class
3: Calculate the probability of fi using the gauss density equation in each class
4: Repeat Step 3 until the probability of all predictor variables (f1,f2,f3,..,fn) has been calculated
5: Calculate the likelihood of each class
6: Get the greatest likelihood

4.2 Support Vector Machines (SVM)

Another approach in supervised learning is to use support vector machines (SVM). By handling discrete and continuous instances, SVM are widely used to detect outliers, perform classification, and perform regression. SVM represent features or occurrences in an n-dimensional space with a defined margin of categories or classes. Using SVM is an excellent choice when working with high-dimensional data. Another advantage of SVM is its memory efficiency. Numerous researchers used SVM, for example, the authors of [16] used this algorithm to predict cardiovascular disease. Their model determined the patient's arterial stiffness by measuring the pulse from the fingertip. The necessary features were retrieved from the reading's waveform. Following that, an SVM classifier was utilized to predict arterial stiffness as low or high. In [17], SVM were used to classify and compare the breathing patterns of patients undergoing weaning trials. The authors of [18] employed this algorithm to classify the heart rate signal. They obtained the reading from three distinct sets of individuals: the young, teenagers, and elders. Twenty individuals were randomly selected from each group, and their heart rates were collected and subsequently categorized using an SVM. Finally, [19] discusses the use of this algorithm in biochemical applications. In this study, a SVM was used to estimate the action potential of the cell membrane.

The pseudocode of SVM [57] is:

Algorithm 2 Pseudo-Code of SVM

Inputs:
Determine the various training and test data
Output:
Determine the calculated accuracy

1: Select the optimal value of cost and gamma for SVM
2: Implement SVM train step for each data point
3: Implement SVM classify for testing data points
4: Repeat Steps 3 and 4 until Stop Condition is met
5: Return calculated accuracy

4.3 Logic based learning

Algorithms based on logic solve problems by sequentially or incrementally applying logical functions. A decision tree is an example of a logic-based learning algorithm. Decision trees are a widely used classification and regression models.

4.3.1 Decision Tree

A decision tree is a logical one composed of nodes and branches. Nodes represent features, while branches represent a value or a condition associated with a node. Classification of the samples is accomplished by sorting them starting with the root node. Sorting is done based on the feature values. At each stage of sorting and selecting the most relevant alternatives, the decision tree makes a determination. It is a straightforward strategy that requires little data preprocessing and is simple to understand. It is, nevertheless, unstable and may result in a complex tree structure. Numerous studies employed decisions trees for classification and regression tasks. The authors of [20] proposed a model for predicting the risk of heart disease based on a patient's health details. A decision tree was used as the basis for the model, and a set of rules was constructed to forecast the risk level. The experimental findings were promising. In study [21], the authors estimated the soil quality using a decision tree model based on the composition of the soil. Study [22] presents a decision tree-based model for Alzheimer's disease prediction. Here, the authors used decision tree induction corresponding to the sample data. At each step/level of the tree, an information gain was used for selecting the feature.

The pseudocode of the Decision Tree algorithm is:

Algorithm 3 Pseudo-Code of Decision Tree Algorithm

1: Place the best attribute of the dataset at the root of the tree
2: Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute
3: Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree

4.4 Instance Based/Lazy Learning

Instance-based or lazy learning postpones generalization until the classification process is completed. As it slows down the process, it is referred to as "lazy learning." Its computational time during the training phase is quite low. In contrast, it is relatively computationally intensive during the classification phase. K-Nearest Neighbor is a widely used instance-based algorithm for classification and regression problems.

4.4.1 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a straightforward algorithm. It is utilized when limited information about the data's distribution exists. KNN classifies new data based on two things: features of said new data and training samples. It stores available data and predicts new data labels based on the similarity measures of the nearest neighbor. It is effective against noisy data and is suitable for training samples taken from real-world situations. However, because the distance to the k nearest neighbor is calculated repetitively for each new data set, the computational cost is quite high. Numerous studies have used KNN to perform classification tasks. It is more typically used for classification tasks than for regression tasks. Study [23] classified heart disease using KNN. In paper [24], the authors suggested a model for offline handwritten digit prediction using KNN. The model was trained and validated using the MNIST digit image dataset. Voting was used to classify the instances. [25] proposed a new KNN model to address the issues associated with evidentiary KNN (EKNN), a KNN extension. Study [26] was proposed by bioscience writers. Here, the authors employed KNN to classify blasts in acute leukemia blood samples. The blasts were classified as either acute lymphocytic leukemia or acute myelogenous leukemia. With an 86% classification accuracy, the results were encouraging. Finally, study [27] proposed a KNN-based approach for the placement of an undergraduate student in an IT business. The classification was binary in nature, as it made use of only two classes (Yes and No).

The pseudocode of the K-Nearest Neighbor algorithm is:

Algorithm 4 Pseudo-Code of K-Nearest Neighbor Algorithm

1: Load the training and test data
2: Choose the value of K
3: Find the Euclidean distance to all training data points
4: Store the Euclidean distances in a list and sort it
5: Choose the first k points
6: Assign a class to the test point based on the majority of classes present in the chosen points
7: Repeat Steps 3, 4, 5, and 6 for each point in test data
8: End

4.5 Deep Learning/Neural Networks

Using Deep learning for classification and regression tasks is another approach in supervised learning. In deep learning, the model comprises many layers, and the model is trained in a layer-by-layer method. Deep learning models have a wide range of applications, from voice recognition to computer vision and natural language processing. By combining several preprocessing techniques and space search optimization, the authors of research [28] suggested a novel framework termed Polynomial Neural Network Classifier (PNNC). In [29], the authors suggested a neural network model for stock price prediction. They named the model FCM (floating centroids Methods), and it reached a high degree of accuracy and optimal operation. In [30], the researchers proposed a neural network-based algorithm for predicting gum disease. Here, a combination of risk factors and symptoms were fed into the model as inputs. Afterward, the hidden layer retrieved features from the given sample and automatically lowered the data's dimensionality. In the end, the output was a binary classification, with 1 indicating the presence of periodontal disease and 0 if gingivitis disease was present. Finally, the authors of [31] employed a multilayered perceptron to forecast heart disease. 13 clinical examples were used to train the model and a prediction was made regarding the existence or absence of cardiac disease. The model worked admirably, with a 98% accuracy rate.

RQ2: What are the widely used evaluation metrics for measuring the performance of the employed supervised learning model?

Several evaluation metrics are used by researchers for the performance evaluation of several classification and regression models. Table 4 presents a list of widely used evaluation metrics.

Evaluation Metric	Definition	Formula
Accuracy	Generally, the accuracy of the prediction model can be defined as the ratio of correct prediction to the total number of input instances	$Acc = \frac{No.ofCorrectPredictions}{TotalNumberofInputInstances}$
Precision	Precision is defined as the number of correct results divided by correctly predicted classes by the prediction model	$P = \frac{TruePositives}{TruePositives + FalsePositives}$
Recall	Recall is defined as number of correct results divided by all relevant samples	$R = \frac{TruePositives}{TruePositives + FalseNegatives}$
F1-Score	F1-Score is the harmonic mean of precision and recall. It shows the robustness of the prediction model	$F_1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$
Mean Absolute Error	It is defined as the average difference between the actual and predicted value	$MAE = \frac{1}{N}\sum_{j=1}^{N} \lvert y_i - \acute{y}_i \rvert$

Continued on next page

Table 4 Continued from previous page

Evaluation Metric	Definition	Formula
Mean Squared Error	Mean Squared Error takes the squared average of the difference between the actual and predicted values	$MSE = \frac{1}{N}\sum_{j=1}^{N}(y - \acute{y}_i)^2$
LogLoss	Logloss also known as logarithmic loss is measured by penalizing the false predictions	$LogLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(P_{ij})$
Area Under Curve (AUC)	AUC shows the ability of the classifier to differentiate between the classes. It is used for binary classification. It depends on true positive rate (TPR) and false positive rate (FPR).	$TPR = \frac{TP}{TP+FN}$ $FPR = \frac{FP}{FP+TN}$

Table 4: Widely Used Performance Metrics

RQ3: What are the merits and demerits of each study in supervised learning?

Table 5 presents several studies, the approach employed and the merits and demerits associated with each of the study.

No	Study Title	Approach	Merit	Demerit	Ref
1	Spam Filtering Using Hybrid Local-Global Naïve Bayes Classifier	Naïve Bayes	A novel learning approach for the classification of messages as spam or legit	Only the independent attributes are taken	[11]
2	Symptom & Risk factor based diagnosis of Gum diseases using Neural Network	Neural Network	Proposed a novel non-invasive approach to diagnose gum disease	Large neural networks require higher computation time	[30]
3	Deep Feature Weighting In Naive Bayes For Chinese Text Classification	Naïve Bayes	A weighted Naïve Bayes model is proposed for Chinese text classification	Error is not mentioned	[12]
4	An Evidential K-Nearest Neighbor Classification Method with Weighted Attributes	KNN	A novel method with weighted features to overcome the issues of EKNN	Did not mention the type of distance and associated attributes that generate better results	[25]
5	Gas Classification Using Binary Decision Tree Classifier	Decision Tree	This study proposes a gas classification model for the electronic nose	Minor change in the tree or the data results into wrong prediction	[20]
6	Cardiovascular Disease Prediction Using Support Vector Machines	SVM	Support vector machine classifies the data into binomial as well as multilevel class	Huge noisy data degrades the performance of SVM	[16]

Continued on next page

Table 5 Continued from previous page

No	Study Title	Approach	Merit	Demerit	Ref
7	Prediction of Heart Disease Using Multilayer Perceptron Neural Network	Neural Network	Achieved an accuracy of 98% in predicting heart disease	NN are not probabilistic	[5]
8	The Application of Decision Tree C4.5 Algorithm to Soil Quality Grade Forecasting Model	Decision Tree	Predicted the soil quality using Decision tree	The training time is costly	[9]
9	A Placement Prediction System Using K-Nearest Neighbors Classifier	KNN	This model predicts the probability of placing an undergrad student in an IT firm	The distance based learning is not defined clearly	[21]
10	Prediction of share price trend using FCM neural network classifier	Neural Network	A method Floating centroids was proposed to predict the share price	High computational time required	[26]
11	ML–XGBoost analysis of language networks to classify patients with epilepsy	XGBoost	Used XGBoost to identify atypical aptterns and classify the participants with focal epilepsy. The results were promising	Required high computational time	[32]
12	An SVM-based ensemble algorithm for breast cancer diagnosis	SVM	Successfully diagnosed breast cancer using SVM with reduced diagnosis variation and improved diagnosis accuracy	Analying large and complex data requires high computational time	[33]
13	A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU	XGBoost	Classified the intrusion detection system dataset. The proposed method outperformed other ML approaches		[34]
14	Applications of SVM Learning in Cancer Genomics	SVM	Comprehend the strengths of SVM in detection of Cancer genomics	Developing new kernel functions is required to deal with the challenges associated with analyzing large and complex data	[35]
15	Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection	SVM RF ELM	Analyze huge amount of data related for intrusion detection. ELM outperformed other ML approaches	Does not provide a description on feature selection and feature transformation techniques in ELM	[36]

Continued on next page

Table 5 Continued from previous page

No	Study Title	Approach	Merit	Demerit	Ref
16	Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm	RF	Proposes a flood disaster resilience evaluation model using RF and overcome the fuzziness of resilience evaluations	Limited indicators	[37]
17	Soil Management Effects on Soil Water Erosion and Runoff in Central Syria A Comparative Evaluation of General Linear Model and Random Forest Regression	RF	Random Forest Regression outperformed General Linear Model in analyzing the most important predictors of soil erosion	For large and complex dataset, due to large number of trees, the RF slows down	[38]
18	Feature Analyses and Modelling of Lithium-ion Batteries Manufacturing based on Random Forest Classification	RF	RF framework attains reliable classification of electrode properties and effectively quantify manufacturing feature importance and correlations	Used only three quantity indicators	[39]
19	AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes	Naïve Bayes RF	Effectively classify disease datasets like diabetes, heart disease, and cancer to check whether the patient is affected or not	Because of the huge amount of data, there is high processing time	[40]
20	COVID-19 World Vaccination ProgressUsing ML Classification Algorithms	Decision Tree KNN Naïve Bayes Random Tree	Decision Tree outperforms other algorithms in terms of time and accuracy	Decision tree can be unstable with a slight change in the data	[41]
21	Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network	Neural Network	Classified COVID-19 chest X-ray images with an accurancy of 93.1%	Requires validating the model with larger dataset and adding explainability component	[42]
22	Differential Deep Convolutional Neural Network Model for Brain Tumor Classification	Neural Network	Proposed model achieved an accuracy of 99.25% infacilitating the automatic classification of brain tumors	An improvement of the parameters is required for differential filter to make the network coverage faster	[43]
23	Risk factors affecting crash injury severity for different groups of e-bike riders: A classification tree-based logistic regression model	Logistic Regression	The logistics analysis was successfully used to indicate the risk factors associated with crash injury severity	Limited data was used for the classification	[44]

Continued on next page

Table 5 Continued from previous page

No	Study Title	Approach	Merit	Demerit	Ref
24	Early diagnosis model of Alzheimer's Disease based on sparse logistic regression	Logistic Regression	SLR improves the classification performance of Alzheimer's disease as compared other classifcal method by reducing the feature dimensions	Only single-mode data is used and LR for multiclass was ignored	[45]
25	Weed Detection Approach Using Feature Extraction and KNN Classification	KNN	KNN classify the weed plant and field crop	Requires high memory and needs to store whole training data	[46]
26	Speech emotion recognition based on SVM and KNN classifications fusion	SVM KNN	To solve the problem of detecting direct feelings of the speaker, a sensation model is presented to classify seven senses	Accuracy depends on data quality, with large data the performance slows down	[47]
27	SVM-Based Traffic Data Classification for Secured IoT-Based Road Signaling System	SVM	SVM detect the anomalies and analying the traffic data pattern. The implementation was done using Raspberry Pi	Time is the only parameter that is considered	[48]
28	Spam email detection using machine learning algorithm	Naïve Bayes	Achieved 88.12% of overall accuracy for spam email detection	Takes all the attributes independantly	[49]
29	Decision tree-based diagnosis of coronary artery disease: CART model	Decision Tree	CART methodology based on Decision tree for CAD classification offered highest diagnosis performance	This classification modelling is highly sensitive in terms of data quality and quantity	[50]
30	A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling	AdaBoost Neural Network	AdaBoost model achieved an high accuracy in classifying the types of soil	11 samples were misclassified	[51]
31	Classification of Skin Diseases Types using Naïve Bayes Classifier based on Local Binary Pattern Features	Naïve Bayes Local Binary Pattern	NB achieved an accuracy of 90% in classifying 9 skin diseases	Combination of these two is only suitable for small dataset	[52]
32	SVM Based Classification and Prediction System for Gastric Cancer Using Dominant Features of Saliva	SVM	Classified the gastric cancer as early gastric cancer and advanced gastric cancer with an accuracy of 97.18%	Complex and noisy data degrades the performance	[53]

Continued on next page

Table 5 Continued from previous page

No	Study Title	Approach	Merit	Demerit	Ref
33	Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis	Neural Network Adaboost	Detect breast cancer with an accuracy of 97.2%	Sensitivity towards outliers and noisy data	[54]
34	Investigating Cardiac Arrhythmia in ECG using Random Forest Classification	Random Forest	Two different arrhythmia conditions in ECG with an accuracy of above 90%	High classification time	[55]

Table 5: Approaches, Merits, and Demerits of Supervised Learning Studies

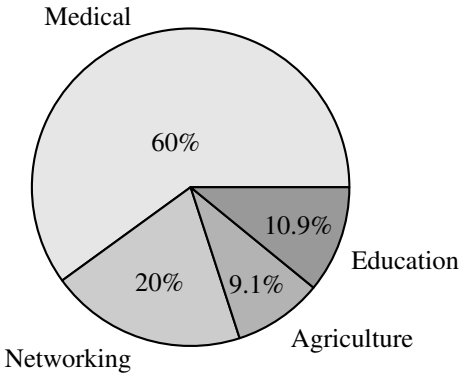

Figure 10: Taxonomy of Papers Based on Research Areas

5 Conclusion

Supervised learning is one of the main categories of ML. It involves training the model on labeled data and testing it on unlabeled data. Additionally, it is divided into classification and regression tasks. Numerous supervised learning algorithms have been proposed throughout the previous decade. Supervised learning is used in a wide variety of applications, from fraud detection to information retrieval, from heart disease diagnosis to cancer detection. This SLR followed Kitchenham's proposed sequence of well-defined phases and is a review of the literature covering supervised learning methodologies and algorithms. It also showcases many performance indicators for supervised learning algorithms. Additionally, it discusses the advantages and disadvantages of many studies. This survey report will assist researchers in determining which supervised learning approach or algorithm to utilize for tackling problems and which area of research requires additional focus.

This survey is limited to widely used supervised learning algorithms and focuses exclusively on research articles published in the last decade and drawn from three databases. In the future, we hope to incorporate other databases, algorithms, and methodologies to improve guidance.

References

[1] James Cussens, "Machine Learning," *IEEE Journal of Computing and Control*, Vol.7, No.4, pp.164-168, 1996.

[2] Muhammad, I., & Yan, Z., "Supervised Machine Learning Approaches A Survey," *ICTACT Journal on Soft Computing*, Vol.5, No.3, 2015.

[3] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, Vol.31, No.3, pp.249-268, 2007.

[4] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning: An Introduction," *Cambridge, MA: MIT Press*, 1998.

[5] Sen, P. C., Hajra, M., & Ghosh, M., "Supervised classification algorithms in Machine Learning: A survey and review," *Emerging technology in modelling and graphics, Springer*, pp.99-111, 2020.

[6] Kadhim, A. I.,"Survey on supervised Machine Learning techniques for automatic text classification," *AI Review*, Vol.52, No.1, pp.273-292, 2019.

[7] Narayanan, U., Unnikrishnan, A., Paul, V., & Joseph, S., "A survey on various supervised classification algorithms," *2017 International Conference on Energy Communication, Data Analytics and Soft Computing (ICECDS), IEEE*, pp.2118-2124, August 2017.

[8] Caruana, R., & Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms," *Proceedings of the 23rd international conference on Machine Learning*, pp.161-168, June 2006.

[9] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Inf. Softw. Technol.*, Vol.51, No.1, pp.7-15, 2009.

[10] Tewari, A. S., Ansari, T. S., & Barman, A. G., "Opinion based book recommendation using naive bayes classifier," *2014 International Conference on Contemporary Computing and Informatics (IC3I), IEEE*, pp.139-144, November 2014.

[11] Solanki, R. K., Verma, K., & Kumar, R., "Spam filtering using hybrid local-global Naive Bayes classifier," *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE*, pp.829-833, August 2015.

[12] Jiang, Q., Wang, W., Han, X., Zhang, S., Wang, X., & Wang, C., "Deep feature weighting in Naive Bayes for Chinese text classification," *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE*, pp.160-164, August 2016.

[13] Bhakre, S. K., & Bang, A., "Emotion recognition on the basis of audio signal using Naive Bayes classifier," *2016 International conference on advances in computing, communications and informatics (ICACCI), IEEE*, pp.2363-2367, September 2016.

[14] Liu, J., Tian, Z., Liu, P., Jiang, J., & Li, Z., "An approach of semantic web service classification based on Naive Bayes," *2016 International Conference on Services Computing (SCC), IEEE*, pp.356-362, June 2016.

[15] Liu, X., Lu, R., Ma, J., Chen, L., & Qin, B., "Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification," *IEEE Journal of Biomedical and Health Informatics*, Vol.20, No.2, pp.655-668, 2015.

[16] Alty, S. R., Millasseau, S. C., Chowienczyc, P. J., & Jakobsson, A., "Cardiovascular disease prediction using support vector machines," *46th Midwest Symposium on Circuits and Systems, IEEE*, Vol.1, pp.376-379, December 2003.

[17] Giraldo, B. F., Garde, A., Arizmendi, C., Jané, R., Diaz, I., & Benito, S., "Support vector machine classification applied on weaning trials patients," *Encyclopedia of Healthcare Information Systems, IGI Global*, pp.1277-1282, 2008.

[18] Kampouraki, A., Nikou, C., & Manis, G., "Robustness of support vector machine-based classification of heart rate signals," *2006 International Conference of the Engineering in Medicine and Biology Society, IEEE*, pp.2159-2162, August 2006.

[19] Seijas, C., Caralli, A., & Villazana, S., "Estimation of action potential of the cellular membrane using support vectors machines," *2006 International Conference of the Engineering in Medicine and Biology Society, IEEE*, pp.4200-4204, August 2006.

[20] Saxena, K., & Sharma, R., "Efficient heart disease prediction system using decision tree," *International Conference on Computing, Communication & Automation, IEEE*, pp.72-77, May 2015.

[21] Dongming, L., Yan, L., Chao, Y., Chaoran, L., Huan, L., & Lijuan, Z., "The application of decision tree C4. 5 algorithm to soil quality grade forecasting model," *2016 First International Conference on Computer Communication and the Internet (ICCCI), IEEE*, pp.552-555, October 2016.

[22] Dana, A. D., & Alashqur, A., "Using decision tree classification to assist in the prediction of Alzheimer's disease," *6th International Conference on Computer Science and Information Technology (CSIT), IEEE*, pp.122-126, March 2014.

[23] Udovychenko, Y., Popov, A., & Chaikovsky, I., "Ischemic heart disease recognition by k-NN classification of current density distribution maps," *35th International Conference on Electronics and Nanotechnology (ELNANO), IEEE*, pp.402-405, April 2015.

[24] Babu, U. R., Venkateswarlu, Y., & Chintha, A. K., "Handwritten digit recognition using K-nearest neighbour classifier," *World Congress on Computing and Communication Technologies, IEEE*, pp.60-65, February 2014.

[25] Jiao, L., Pan, Q., Feng, X., & Yang, F., "An evidential k-nearest neighbor classification method with weighted attributes," *Proceedings of the 16th International Conference on Information Fusion, IEEE*, pp.145-150, July 2013.

[26] Supardi, N. Z., Mashor, M. Y., Harun, N. H., Bakri, F. A., & Hassan, R., "Classification of blasts in acute leukemia blood samples using k-nearest neighbour," *8th International Colloquium on Signal Processing and its Applications, IEEE*, pp.461-465, March 2012.

[27] Giri, A., Bhagavath, M. V. V., Pruthvi, B., & Dubey, N., "A placement prediction system using k-nearest neighbors classifier," *Second International Conference on Cognitive Computing and Information Processing (CCIP), IEEE*, pp.1-4, August 2016.

[28] Huang, W., Oh, S. K., & Pedrycz, W., "Polynomial neural network classifiers based on data preprocessing and space search optimization," *Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS), IEEE*, pp.769-773, August 2016.

[29] Liu, S., Yang, B., Wang, L., Zhao, X., Zhou, J., & Guo, J., "Prediction of share price trend using FCM neural network classifier," *3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS), IEEE*, pp.81-86, August 2016.

[30] Thakur, A., Guleria, P., & Bansal, N., "Symptom & risk factor based diagnosis of Gum diseases using neural network," *6th International Conference-Cloud System and Big Data Engineering (Confluence), IEEE*, pp.101-104, January 2016.

[31] Sonawane, J. S., & Patil, D. R., "Prediction of heart disease using multilayer perceptron neural network," *International conference on information communication and embedded systems (ICICES), IEEE*, pp.1-6, February 2014.

[32] Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M., "Machine Learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain informatics*, Vol.4, No.3, pp.159-169, 2017.

[33] Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S., "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *European Journal of Operational Research*, Vol.267, No.2, pp.687-699, 2018.

[34] Bhattacharya, S., Maddikunta, P. K. R., Kaluri, R., Singh, S., Gadekallu, T. R., Alazab, M., & Tariq, U., "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU," *Electronics*, Vol.9, No.2, pp.219, 2020.

[35] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W., "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics-Proteomics*, Vol.15, No.1, pp.41-51, 2018.

[36] Kavin, S., Mohan, S. K., Karthick, V. I., & Sudar, K. M., "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *Special Section on Survivability Strategies for Emerging Wireless Networks, IEEE*, 2018.

[37] DongSheng Liu, Zhongrui Fan, Q. Fu, Mo Li, M. A. Faiz, Shoaib Ali, Tianxiao Li, L. Zhang, Muhammad Imran Khan, "Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm," *Journal of Cleaner Production*, Vol.250, pp.119468, 2020.

[38] Safwan Mohammed, Ali Al-Ebraheem, Imre J Holb, Karam Alsafadi, Mohammad Dikkeh, Quoc Bao Pham, Nguyen Thi Thuy Linh, Szilard Szabo, "Soil management effects on soil water erosion and runoff in central Syria—A comparative evaluation of general linear model and random forest regression," *Water*, Vol.12, No.9, pp.2529, September 2020.

[39] Liu, K., Hu, X., Zhou, H., Tong, L., Widanalage, D., & Marco, J., "Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification," *IEEE/ASME Transactions on Mechatronics*, 2021.

[40] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y., "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Journal of Supercomputing*, Vol.77, No.5, pp.5198-5219, 2021.

[41] Abdulkareem, N. M., Abdulazeez, A. M., Zeebaree, D. Q., & Hasan, D. A., "COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms," *Qubahan Academic Journal*, Vol.1, No.2, pp.100-105, 2021.

[42] Abbas, A., Abdelsamea, M. M., & Gaber, M. M., "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," *Applied Intelligence*, Vol.51, No.2, pp.854-864, 2021.

[43] Abd El Kader, I., Xu, G., Shuai, Z., Saminu, S., Javaid, I., & Salim Ahmad, I., "Differential deep convolutional neural network model for brain tumor classification," *Brain Sciences*, Vol.11, No.3, pp.352, 2021.

[44] Wang, Z., Huang, S., Wang, J., Sulaj, D., Hao, W., & Kuang, A., "Risk factors affecting crash injury severity for different groups of e-bike riders: A classification tree-based logistic regression model," *Journal of safety research*, Vol.76, pp.176-183, 2021.

[45] Xiao, R., Cui, X., Qiao, H., Zheng, X., & Zhang, Y., "Early diagnosis model of Alzheimer's Disease based on sparse logistic regression," *Multimedia Tools and Applications*, Vol.80, No.3, pp.3969-3980, 2021.

[46] Khurana, G., & Bawa, N. K., "Weed Detection Approach Using Feature Extraction and KNN Classification," *Advances in Electromechanical Technologies, Springer*, pp.671-679, 2021.

[47] Al Dujaili, M. J., Ebrahimi-Moghadam, A., & Fatlawi, A., "Speech emotion recognition based on SVM and KNN classifications fusion," *International Journal of Electrical and Computer Engineering*, Vol.11, No.2, pp.1259, 2021.

[48] Sankaranarayanan, S., & Mookherji, S., "SVM-based traffic data classification for secured IoT-based road signaling system," *Research Anthology on AI Applications in Security, IGI Global*, pp.1003-103, 2021.

[49] Nayak, R., Jiwani, S. A., & Rajitha, B., "Spam email detection using Machine Learning algorithm" *Materials Today: Proceedings*, 2021.

[50] Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A., "Decision tree-based diagnosis of coronary artery disease: CART model" *Computer methods and programs in biomedicine*, Vol.192, pp.105400, 2020.

[51] Binh Thai Pham, Manh Duc Nguyen, Trung Nguyen-Thoi, Lanh Si Ho, Mohammadreza Koopialipoor, Nguyen Kim Quoc, Danial Jahed Armaghani, Hiep Van Le, "A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling," *Transportation Geotechnics*, Vol.27, pp.100508, 2021.

[52] Putri, H. S. K. A., Sari, C. A., & Rachmawanto, E. H., "Classification of Skin Diseases Types using Naïve Bayes Classifier based on Local Binary Pattern Features," *International Seminar on Application for Technology of Information and Communication (iSemantic), IEEE*, pp.61-66, September 2020.

[53] Aslam, M. A., Xue, C., Wang, K., Chen, Y., Zhang, A., Cai, W., ... & Cui, D., "SVM based classification and prediction system for gastric cancer using dominant features of saliva," *Nano Biomed Eng*, Vol.12, No.1, pp.1-13, 2020.

[54] Zheng, J., Lin, D., Gao, Z., Wang, S., He, M., & Fan, J., "Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis," *IEEE Access*, Vol.8, pp.96946-96954, 2020.

[55] Kumar, R. G., & Kumaraswamy, Y. S., "Investigating cardiac arrhythmia in ECG using random forest classification," *Intl. J. Comput. Appl*, Vol.37, No.4, pp.31-34, 2012.

[56] Muhammad Firman Saputra, Triyanna Widiyaningtyas, Aji Wibawa, "Illiteracy Classification Using K Means-Naïve Bayes Algorithm," *International Journal on Informatics Visualization*, Vol.2, No.153, 10.30630/joiv.2.3.129, 2018.

[57] Houssein, E.H., Hosney, M.E., Elhoseny, M. et al., "Hybrid Harris hawks optimization with cuckoo search for drug design and discovery in chemoinformatics," *Sci Rep*, Vol. 10, pp.14439, 2020.