Early Stage Diabetes Prediction by Approach Using Machine Learning Techniques

Muhammad Zarar
Wuhan University
Yulin Wang (✉ yulinwang@whu.edu.cn)
Wuhan University

Research Article

Keywords: ANN, MLA, AI, PIMA, Kaggle, Diabetes Prediction, Machine Learning.

Posted Date: July 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3145599/v1

Abstract

Diabetes is the most viral and chronic disease throughout the world. A large number of people are affected by this chronic disease. Early detection of diabetes in a patient is crucial for ensuring a good quality of life. Machine learning techniques or Data Mining Techniques are playing a significant role in today's life to detect diabetes and improve performance to make further accurate predictions. The aim of this research is diabetes prediction with the approach of machine learning techniques. In this technical approach, we have taken two data sets Pi-ma Indian diabetes data set and the Kaggle diabetes data set, and proposed a model for diabetes prediction. We have used four different machine learning algorithms such as Support Vector Machine, Decision Forest, Linear Regression, and Artificial Neural Network. In these machine learning algorithms, ANN gives the best prediction performance where the highest accuracy is 98.8% so, it could be used as an alternative method to support predict diabetes complication diseases at an initial stage. Further, this work can be extended to find how likely non-diabetic people can have diabetes in the next few years and also, this predicted model can be used for imaging processing in the future to find diabetes for the prediction of diabetic and non-diabetic.

1 Introduction

Researchers define AI as a broad application area of electrical engineering and computer science technology. AI is an amalgamation of two terms: "artificial," which refers to things that are not natural or man-made, and "intelligence," which denotes the ability to learn, solve problems, and make decisions. In essence, AI can be defined as the automation of intelligent behavior and the capacity of computer programs and machines to think and learn. AI encompasses various aspects such as thinking, learning, problem-solving, acting rationally, and mimicking human thought processes [1].One crucial aspect of AI is the simulation of human behavior and intelligence by machines. By leveraging advanced algorithms and computational power, AI systems can analyze vast amounts of data, recognize patterns, and make informed predictions or decisions. The objective of AI is to develop adaptive tools in a dynamic environment, utilizing insights such as the ability to avoid improbable solutions. For instance, in game playing and puzzle solving, AI algorithms can analyze the game's rules and optimize strategies to defeat human opponents or achieve the best possible outcome. In computer vision, AI enables machines to perceive and interpret visual data, opening doors to applications such as facial recognition, object detection, and autonomous vehicles. In the field of robotics, AI empowers machines to perform complex tasks with precision and adaptability, enhancing efficiency in manufacturing, exploration, and healthcare domains.

AI finds applications in diverse domains, including expert systems that emulate human expertise in specialized areas, automotive and heavy industries that employ AI for automation and optimization, healthcare that utilizes AI for diagnostics and treatment recommendations, and data mining that harnesses AI algorithms to extract valuable insights from large datasets. The integration of AI into various sectors has yielded significant advancements and transformative impacts, revolutionizing the way we live, work, and interact with technology [2]. In this study, we will be discussing the field of data mining, which represents a significant area with in the domain of Artificial Intelligence. Data mining pertains to the extraction of valuable insights and knowledge from large datasets using intelligent machines. AI term was first introduced as an academic direction in 1956 by American PC researcher John McCarthy John Mc Carty [3], sorted out the Dartmouth conference, at which the term “Man-made brainpower” was first embraced. The motivation part of data mining, earlier people suffered to collect a lot of data, so the people suggested various methods for collecting the data, storing and managing the data where behind the data base management system [4]. Scientists use information mining approaches like multi-dimensional databases, machine learning, delicate registering, information representation and statistics, healthcare fields, and many others. It is also known as “Information Mining” or “Discovery of the knowledge” and “Knowledge Extraction about the data”.

Diabetic mellitus is a worldwide dangerous disease. It could also affect other parts of the human body, simply, it is a risky disease, and millions of people have suffered from this chronic disease. The prediction of this chronical deases are very important in early stage of life. Before this study many diabetes prediction models were used but on the basis of result their accuracy rate was very low and could not give better prediction. Though in real fact the patient essential predict either to be in diabetic group or non-diabetic group. This blunder in diagnosis may lead to needless treatments or no needless at all when required [5]. In order to avoid or reduce brutality of such impact, there is a need to create a model using machine learning algorithm and data mining techniques which we provide accurate results and shrink human efforts from the old prediction model we improved accuracy rate and give the better result of the patient to predict diabetic or non-diabetic, using machine learning algorithm and data mining techniques which will provide accurate results and reduce human efforts. We used different machine learning algorithms but among these ANN gave us best result. The proposed model of this research is to predict the diabetes of the patient. We also improve the accuracy rate up to 98.8% percent by ANN and it is still high rate.

2 Literature Review

We proposed a model to predict Type-2 diabetes using machine learning algorithm (MLA). Type-2 diabetes is chronic stage of the diabetic, which is insulin depended [6]. Dataset used in this paper was collected from an online survey. Total 952 participants were involved in this survey. Among these 372 were Females and 580 are males. Age of each participant was above 18 and another dataset PIDD was also used for the comparison. Using six different MLA, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes and Random Forest. The dataset was split into 75:25 for testing and training by bi-classification method, In the resultant highest accuracy was 94.10% from the RF. Amelie Viloria et al [7].The motive of this study was to propose a methodology for the diabetes diagnostic disease using SVM for the prediction of diabetes diagnosis, using major instances of the human body like DM, Age, BMI, BG for the diagnosis of the diabetes disease. The dataset was collected from the Colombia hospital having 500 patients record including males and females. The evaluation of this dataset output variables are three, Yes diabetes, No diabetes and Free disposition to diabetes. The 80% dataset used to train non-linear SVM for classification in a patient and 20% dataset used for validation.

10-cross validation methods were used to authenticate computational model and confusion matrix method used for the measurements. The highest accuracy obtained from SVM was 92.2% recorded.

Deepti Sisodia and Dilip Sing Sisidia projected a model to predict of diabetes using classification algorithms [8] to give maximum accuracy of most efficiently in result. There are three machine learning classification algorithms like Decision Tree (DT), Support Vector Machine (SVM) and Naıve Bayes (NB) were used to classify the data and find the result. The data was collected from Pima Indians Diabetes Dataset (PIDD) were includes the strength of 786 Instants and 8 attributes ware used. The results of two types of data, one is positive show “1” and second is negative while show “0” using WEKA tool for analysis of data. All the algorithms results evaluate for pre-processing and applied F-measures and recalled operations. The maximum accuracy is 76.30% from Naıve Byes among them. Swapna G et al [9]. In this paper projected a methodology for Diabetes Detection using Deep Learning Algorithms (DLA). This model used 3 different DLA, its name was Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and the combinations of these two CNN-LSTM which was called hybrid (CNN). The input data of this algorithms was Heart Rate Variability (HRV) and was derived from the Electro Cardiogram (ECG) .The data evaluation for the classification using Support Vector Machine (SVM) techniques. In the result gain best performance of CNN and LSTM-CNN is 0.03% and 0.06% before applied SVM. The maximum accuracy of this classification system to diagnose diabetic using ECG waves is 95.7%. Soman KP etal. The aim of this paper was to proposed architecture to Detect of Diabetes Automated using CNN and LSTM Network and Heart Rate Signals (HRV) [10]. For the classification of data extraction were used Deep Learning Techniques (DPL). The input data was split into two types which one was used for testing and the other was used for training. The 5-Fold Cross Validation Method was used for confusion matrix. The accuracy of cross validation was found 93.6% and gave detail accuracy 95.1%. In the resultant output value 0 means diabetes and 1 mean no diabetes. Aishwarya Mujumdar and Dr. Vaidehi V are the authors of this paper were proposed diabetes prediction model by using MLA. DPM had five different type of modules which is Dataset Collection, Data Pre-Processing, Clustering, Build Model, and Evaluation [11]. The data collection of the data set was Pima Indian Diabetic Dataset which is 800 recorded people and 8 attributes. Data pre-processing cleared all missing values as after this phase clustering modules were used K-means clustering algorithms on data set. In this model algorithms includes Support Vector Classifier, AdaBoost Classifier, Decision Trees Classifier, Random Forest Classifier, Extra Tree Classifier, Logistic Regression, K-Nearest Neighbor, Gaussian Na¨ıve Bayes, Gradient Boost classifier. In the last evaluation of the classification accuracy was confusion matrix and f1-score. The resultant maximum accuracy was 98.8% gain form the AdaBoost Classifier.

Han Wu et al [12]. Motive of this journal paper was the prediction of type-2 diabetes mellitus using data mining techniques. T2DM is that type of diabetes patient which is non-insulin dependent. This study prepared a novel model for the prediction of T2DM using two mining classification algorithms, Logistic Regression and K-means Cluster. The Pima Indian Diabetes data set (PIDD) used for the classification and prediction [13]. The data set is containing 768 recorded patients were 268 positive tested and 500 are negative tested instances. WEKA toolkit was used for analysis of data for pre-processing. The experimental process K-fold validation method used for the verify performance of a model. The evaluation of the models collect data through online questionnaire which contained 384 instances which was divided into two groups one is 68 positive and 316 negatives. The model obtained 3.04% above results with other researcher results. The maximum accuracy obtained from this model was 94%. Quan Zou et al [14]. The intention of this study predicting diabetes mellitus with machine learning techniques. Two data sets were used in this paper, the data was collected from hospital in Luzhou, China, this data set is split into two parts: the healthy people and diabetes affected people and another data set was used PIDD there have all patients are female were age is 21-year-old. Among these data set samples were randomly selected [15]. Three different classification algorithms Decision Tree, Random Forest and Neural Network were applied. The results of algorithms were compared collected data set from hospital in Luzhou China and PIMDD. The maximum accuracy in results show that the prediction of DM could be reached highest accuracy 0.8084 from RF.

Tanha Mehboob Alam et al. The objective of this research paper was preparing a model for early prediction of diabetes using data mining techniques [16]. Data set used in this paper, was originally taken from the National Institute of Diabetes, Digestive and Kidney disease (publicly available). In this model inconsistency, noise and missing values of the data set were removed by using three methodologies, Data preprocessing, Data cleaning and Data reduction. Three models were used in this paper which was Artificial Neural Network (ANN), Random Forest (RF) and K-means clustering. The maximum prediction accuracy was 75.7% derived from ANN. Changsheng Zhu et al [17]. The goal of this study was to design data mining-based model for early diagnosis and prediction of diabetes using PCA (Principal Component Analysis) and K-means clustering techniques. The data obtained from PIDD data set was of total 768 samples female patient 500 tested was negative object and 268 is positive tested object. Total 8 attributes of the data set were one class is label. Applying pre-processing method for removing noisy, in consistence and missing values of the data set. Two algorithms K-means clustering and LR applying for classifications [18]. The best accuracy of this model was 89.0%.Tejas and M. Chawan focuses on the topic of diabetes, a chronic disease with a significant impact on global healthcare. It highlights the alarming statistics provided by the International Diabetes Federation, emphasizing the growing number of individuals living with diabetes worldwide [19]. The review acknowledges the challenges in early prediction of diabetes due to its complex interdependence on various factors and its adverse effects on multiple organs. It then introduces the application of data science methods, particularly machine learning, in the medical field to improve predictions and diagnostic capabilities. The review outlines the aim of the project, which is to develop a system that combines different machine learning techniques, including SVM, Logistic regression, and ANN [20], to achieve accurate early prediction of diabetes. The goal is to contribute to the development of effective techniques for earlier detection of this disease. Authors Jobeda Jamal Khanam and Simon Y. Foo conducted research on diabetes prediction, utilizing data mining, machine learning algorithms, and neural network method [21]. The study focused on the Pima Indian Diabetes dataset obtained from the UCI Machine Learning Repository, which included information on 768 patients and nine attributes. Applying seven machine learning algorithms, the researchers found that Logistic Regression (LR) and Support Vector Machine (SVM) performed well in predicting diabetes. Additionally, a neural network model with two hidden layers achieved an accuracy of 88.6% after testing different epochs. Diabetes is a pervasive global health concern that affects individuals worldwide, resulting in elevated blood sugar levels and various complications. Despite efforts to develop accurate diabetes prediction models, researchers face significant challenges due to the scarcity of suitable data sets and prediction approaches [22] to address these issues, this study employs big data analytics and machine learning methods to explore predictive analytics in healthcare.

The primary objective is to construct an intelligent framework specifically tailored for diabetes prediction. By utilizing decision tree-based random forest and support vector machine models, the researchers propose the Intelligent Diabetes Mellitus Prediction Framework (IDMPF) [23]. This framework is the result of a comprehensive literature review, offering promising results with an accuracy rate of 83%. The findings of this study have implications for healthcare professionals, stakeholders, and researchers involved in diabetes prediction research and development [24].

Table 1
Home ground of Relevant Literatures

Years	Techniques	Datasets	Results
2018	DT, SVM, NB	PIDD	76.30%
2018	RNN, LST, CNN, CNN-LSTM, SVM	ECG Signals	95.7%
2018	CNN, LST, CNN-LSTM	ECG Signals	95.1%
2018	K-means, LR	Private + PIDD	95.42%
2018	DT, RF, NN	Private + PIDD	0.8084
2019	SVM, ABC, DT, RF, ET, LR, KNN, GNB, GB	Private + PIDD	98%
2019	ANN, RF, K-means	NIDD	75.7%
2019	PCA, K-means, LR	PIDD	97.40%
2020	LR, KNN, SVM, NB, DT, RF	Private + PIDD	94.4%
2020	SVM	Private	95.36%
2021	LR, SVM	PIDD	88.6%
2022	RF, SVM, KNN, LR	PIDD	86%

These findings tabele.1 contribute to a broader understanding of the subject matter and provide a foundation for further analysis and exploration. By synthesizing and comparing the results from different studies, important patterns, trends, and correlations can be identified. This allows for a comprehensive view of the topic, enabling researchers to draw meaningful conclusions and make informed decisions based on the collective knowledge gained from these diverse sources.

3 Research Model, Design and Analytical Approach

The methodology employed in this research focuses on predicting early stage diabetes through the utilization of machine learning techniques. To achieve this, a comprehensive dataset consisting of relevant health parameters and clinical measurements will be collected from a diverse group of individuals. The collected data will then undergo a rigorous preprocessing phase, including data cleaning, feature selection, and normalization, to ensure its quality and suitability for analysis. Subsequently, a range of machine learning techniques, such as logistic regression, decision trees, and support vector machines, will be applied to the preprocessed data to develop research or predictive models. The performance of these models will be evaluated using appropriate evaluation metrics, such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve [25] by following this methodology, valuable insights can be gained regarding the early detection and prediction of diabetes, potentially leading to timely inter venations and improved healthcare outcomes.

3.1 Data Source Description

In this diabetes prediction research study, two distinct datasets were utilized to analyze and predict the likelihood of diabetes occurrence. The datasets, obtained from a publicly available health databases centers, consists of anonymized medical records of individuals patients. It encompasses a wide range of features, including demographic information, clinical measurements, and historical medical records. This dataset enables the exploration of various risk factors and their association with diabetes development [26] by combining these two datasets, we aim to develop a comprehensive predictive model that takes into account both clinical and lifestyle factors for accurate diabetes risk assessment especially in early stage. The datasets provide a robust foundation for examining the relationships between different variables and developing an effective prediction model to aid in early detection and intervention strategies for diabetes management [27].

3.1.1 Kaggale Diabetes Dataset:

The main motive of this dataset is used for prediction of diabetes. It is substantial from the National Institute of Diabetes Diseases (NIDD) [28]. This dataset consists of 2000 records, 1316 instances are negative test recorded which having value is “0” and 684 instances are positive tested record which having value is “1”. Each instance having 8 attributes (9 = 8 + 1 class attribute) having all numerical numbers and representing the data in the dataset is personal health data as well as results from medical examinations. These attributes encompass various factors associated with diabetes, allowing for a comprehensive exploration of the disease. The dataset offers valuable information on key indicators, such as glucose levels, blood pressure, skin thickness, pregnancies, body mass index (BMI), insulin resistance, age and diabetes pedigree function.

By leveraging this dataset, our research aims to uncover patterns, correlations, and predictive insights that can contribute to a better understanding of diabetes and aid in the development of effective diagnostic and preventive measures. A graphical representation of the dataset is showing above figure.

3.1.2 Pima Indians Diabetes Dataset:

This dataset is derived from NIDD which is publicly available at UCI machine repository. The impartial of this dataset is to predict diagnosis whether a patient has diabetes or not diabetes, based on convinced diagnostic measurements which include in dataset. This dataset, collected from the Pima Indian population, provides a valuable resource for examining the relationships between these attributes and the occurrence of diabetes [29]. It consists of 768 of the female patient records and encompasses eight key attributes and having 9 (9 = 8 + 1 class attribute) same attributes as above collection of data were the resulting total of 268 tested positive instances and 500 tested negative instances.

Analyzing this dataset, our research aims to gain insights into the risk factors and potential predictors of diabetes and ultimately contributing to improved understanding and prevention strategies for this condition.

3.2 Data Pre-Processing:

Pre-processing method is one of the superlative data mining technique in the field of diabetes prediction using machine learning techniques, effective data preprocessing methods are crucial for optimal data mining outcomes. These techniques encompass essential tasks such as data normalization, data integration, data transformation, and data reduction. By employing these pre-processing methods, the dataset's integrity is ensured, preventing erroneous or unreliable predictions [30]. Initially, a comprehensive attribute analysis was conducted to evaluate the dataset's suitability for diabetes prediction and diagnosis. To facilitate a streamlined analysis process, numeric attributes were converted into nominal attributes, effectively reducing the complexity associated with dataset examination [31]. This preprocessing step was particularly vital as low-quality or incomplete data can yield inaccurate or subpar prediction results.

Furthermore, to enhance the dataset's uniformity and simplify subsequent calculations, an unconfirmed standardized filter attribute was applied. This filter attribute standardized all the data within the range of [0, 1]. The standardization formula employed a mean or average value ('nan') and the normal deviation ('t') for each variable ('V'). The resulting normalized value ('N') provided a consistent framework for further computations, significantly improving computational efficiency [32].

$$N = \frac{V - nan}{t}$$

By implementing these meticulous data preprocessing steps, the research aims to mitigate data-related complications, optimize predictive accuracy, and expedite the overall operational speed.

3.3 Model Architecture and Design:

The research presents a sophisticated model design and architecture tailored for early-stage diabetes prediction. The model will undergo comprehensive training utilizing a carefully curated dataset exclusively designed to address the intricacies of early detection. During the training process, a joint optimization of the feature extraction and classification components will be performed. This simultaneous parameter optimization empowers the model to acquire a deep understanding of the complex patterns and interdependencies inherent in the data, thereby facilitating accurate predictions on previously unseen instances.

It is worth highlighting that the selection of the specific model architecture and techniques will be contingent upon the unique characteristics of the dataset and the specific research objectives. The proposed architecture and associated techniques will be thoughtfully adapted to accommodate the distinctive challenges and goals of the study, ensuring an optimal fit between the research requirements and the chosen model design as shown in the above figure.4.

3.3.1 Testing and Training

The training and testing process plays a crucial role in the development and evaluation of the predictive models for early stage diabetes. During the testing and training phases, the algorithms acquire knowledge from the training data and subsequently generate predictions or make informed decisions [33]. To ensure an accurate assessment of algorithm performance, the datasets are divided into two subsets: a training set and a testing or validation set. The training set constitutes 80% of the original dataset, providing a robust foundation for algorithm training, while the remaining 20% is dedicated to testing and validation.

The evaluation of the dataset during both training and testing, the Anaconda toolkit was employed. This comprehensive toolkit offers a wide range of analytical tools and resources, facilitating efficient dataset evaluation in this particular methodology. Below code demonstrates the dataset splitting process.

```
X train, X test, y train, y test = train test split

(X, Y, test size = 0.2)
```

Above code illustrates the utilization of the train-test-split function, a fundamental component of the evaluation process. It divides the dataset into X-train and X-test, representing the feature matrices, as well as y-train and y-test, which correspond to the respective target variables. This partitioning of the dataset ensures that a dedicated portion is reserved for unbiased testing and validation.

3.4 Emerging Classification Approach

The Emerging Classification Approach section of this research explores the application of four distinct machine learning techniques: Random Forest, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN). These techniques have garnered considerable attention in the field of diabetes prediction and exhibit promising potential for accurate classification by incorporating these diverse machine learning techniques [34], this research aims to explore the effectiveness of each approach and identify the most suitable method for diabetes prediction. This comprehensive analysis contributes to the advancement of classification methodologies in the context of diabetes research. Details of these four methods are describe below.

3.4.1 Random Forest Classifier:

The Random Forest classifier, an effective collective learning method, proves invaluable in predicting diabetes mellitus. This flexible and versatile machine learning algorithm combines multiple decision trees to yield robust and reliable outcomes. The Random Forest method employs an aggregate decision-making approach, considering the individual decisions of the constituent trees in a sequential manner. This sequential aggregation process ensures the dataset properly, secondly, normalize the dataset, thirdly is ensemble the decision tree, fourthly, aggression means that take majority decisions and it predict the outcomes. The result is a more accurate and dependable prediction, as depicted in the figure below.

Random forest classifier, use the mean squad error equation (MSEE). This is used for how much data parts form each nodes of the decision of your forest analysis:

$$RFMSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

The performing of RF classification data we should use the Gini Index for evaluation that how much nodes on the data points and Entropy.

$$Gini = 1 - \sum_{i=0}^{C}(P_i)^2$$

$$Entropy = 1 - \sum_{i=0}^{C}P_i * log_2\left(P_i\right)$$

‘Pi’ presents the relative frequency of the class and ‘C’ is the number of the classes. Entropy used for the probability of certain results and log function were used for calculating mathematical intensive.

3.4.2 Decision Tree Classifier:

In the realm of early-stage diabetes detection, the Decision Tree Classifier emerges as a promising tool for developing predictive models. Leveraging a collected dataset comprising crucial health parameters and clinical measurements [35], this algorithm enables the construction of a decision tree model. Through the utilization of various input features, the model learns the underlying patterns and relationships within the data, thereby facilitating precise predictions concerning the probability of an individual having early-stage diabetes.

This algorithmic approach offers interpretability, allowing for the identification of significant predictors and providing a clear understanding of the decision-making process. The derived decision tree model serves as a valuable tool for risk assessment and early intervention, enabling timely and targeted preventive measures for individuals at higher risk of developing diabetes. Implementing the Decision Tree

Classifier in this research further contributes to the ever-evolving landscape of early-stage diabetes detection, paving the way for more accurate and personalized healthcare strategies.

3.4.3 Support Vector Machine:

SVMs are a popular machine learning algorithm used for both classification and regression tasks. Particularly, SVM classifiers excel in analyzing and predicting numerical data in binary classification scenarios. In the context of this study, SVMs offer a powerful framework for building a robust predictive model. SVMs provide the means to identify and exploit significant features, facilitating precise predictions and aiding in the understanding of the underlying factors contributing to early-stage diabetes.

The primary objective of SVMs is to classify data points by finding the hyperplane that maximally separates them in a multidimensional space, as depicted in the figure below [36]. This hyperplane acts as a decision boundary, effectively distinguishing between different classes of data points. Hyperplane is an area to classify the data point in domain.

3.4.4 Artificial Neural Network:

Artificial Neural Networks (ANN) play a crucial role as an information processing model inspired by the biological neuron system. By mathematically modeling neural processes, It is neural implemented to model. ANNs offer a powerful tool for pattern recognition and data classification. In my study, I aim to explore the effectiveness of ANN in analyzing complex datasets and predicting disease outcomes. Drawing inspiration from the billions of interconnected neurons in the human brain, ANNs consist of multiple nodes that emulate artificial neurons. These nodes, or artificial neurons, are connected to each other, forming a network that facilitates the collection and processing of data. Just as information is transmitted through the connections between neurons in the brain, Information is recognized by dendrites which is taken as stimuli, the input data in my study is transmitted through these connections in the ANN, with each connection assigned a specific weight to influence the information flow. By harnessing the computational power and interconnectedness of ANNs, I anticipate uncovering valuable insights and achieving accurate predictions in diabetes prediction model [37].

The model of ANN is consisting into layers of multiple nodes, the data travels from first layer this is called input layer, and after passing through middle layers or hidden layers it reaches the output layer, every layer transforms the data into some related information and to end gives the desired result.

$$f\left(\mathrm{b}+\sum_{i=0}^{n}(i_0.\,w_0)+(i_1.\,w_1)+\ldots(i_n.\,w_n)\right)$$

Above mathematical equation there “f” is the activation function (output, f (I) = I) were applied to input weight, ‘w’ is consistent weights and ‘x’ is inputs to given function.

4 Experimental Findings and Discussion

Encompasses the successful collection of final results, aimed at predicting diabetes mellitus using various machine learning algorithms. Through the application of different algorithms, including Random Forest, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN), the research explored the predictive capabilities of each method. Two distinct diabetes datasets, namely PIMA and Kaggle, were utilized to evaluate the performance of these four supervised learning approaches. Applying the machine learning algorithms to both datasets, diverse accuracies were obtained, signifying the effectiveness of each technique in classifying patients as diabetic or non-diabetic. Notably, the Artificial Neural Network (ANN) yielded the highest accuracy of 98%, as depicted in table.

Table 2
Accuracy outcomes Analyze

Method	Dataset	Accuracy	F1 Measure	Precision	Recall
Random Forest	Kaggle	77.5%	66%	85%	54%
	PIDD	77.9%	69%	81%	60%
Decision Tree	Kaggle	95.5%	85%	70%	72%
	PIDD	92%	80%	85%	73%
SVM	Kaggle	97.2%	95%	96%	92%
	PIDD	74%	75%	76%	73%
ANN	Kaggle	98.9%	96%	92%	85%
	PIDD	98%	95%	94%	80%

These findings demonstrate the potential of machine learning algorithms in accurately predicting diabetes mellitus. The high accuracy achieved by the ANN model highlights its efficacy in discerning intricate patterns and relationships within the datasets. The results also underline the importance of employing supervised learning techniques in diabetes diagnosis, enabling healthcare professionals to make informed decisions and provide targeted interventions. The varying accuracies obtained from the different algorithms indicate the significance of selecting an appropriate approach based on the specific dataset and research objectives. This underscores the need for a comprehensive evaluation and comparison of multiple algorithms to identify the most suitable method for diabetes prediction.

4.1 Evaluation and Result Analysis

The final step in assessing the proposed diabetics' prediction model within this research. This crucial phase involves the comprehensive evaluation of various methods utilized for diabetes prediction, employing a range of accuracy measurements to gauge their performance. These accuracy measurements during the evaluation phase, can gain a comprehensive understanding of the performance of the diabetics' prediction model. The combination of classification accuracy, provides a robust framework [39] for assessing the effectiveness and reliability of the model in predicting diabetes. The following key accuracy measurements, including classification accuracy, confusion matrix,f-1 score, recall, and precision, are defined below and serve as essential evaluation metrics. This metric provides insights into the model's ability to accurately identify and capture positive instances within the dataset [40].

4.1.1 Classification Accuracy

Classification accuracy is defined the numbers of the correct predictions are divided by the total number of inputs samples or total number of predictions made. It is mathematically given as:

$$CA = \frac{(NCP)}{(TNI)}$$

Allowing to this experiment the performance of artificial neural network is capable with highest accuracy is 98.9%. Decision tree is much closed result to ANN, having 92.7% accuracy.

4.1.2 Confusion Matrix

Confusion matrix is doing well performance for the binary classification methods; we are also using binary classification in this research. Confusion matrix is giving the result as form of matrix, where describe the full performance of the proposed model. It gives us 4 values and two classes (actual class and predict class) in output.

The mathematical representation of the average accuracy of confusion matrix shown below, where ‘N’ is the total number of inputs:

$$Accuracy = \frac{TP + FN}{N}$$

4.1.3 F1 Measure

F1-score is also called F-measures. It is used for the measure of test’s accuracy and identifying the number of true and positive of the precision and recall. It is the harmonic means value of the precision and recall. In this experiment the highest fi values of AAN is 96%.

Mathematically represent as:

$$FM = 2 \times \frac{precssion * recall}{precission - recall}$$

4.1.4 Precision:

Precision define is the fraction of true positive values among number of positive values predicted by the classifier. It is expressed as:

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

4.1.5 Recall:

Recall, also referred to as sensitivity or true positive rate, and represents the ratio of correctly predicted positive outcomes to the total number of samples that are actually positive.

Mathematically, it can be expressed as:

$$Precision = \frac{(TP)}{(TP) + (FN)}$$

5 Conclusion

In this experimental analysis, the research aimed to predict diabetes in the early stages using two distinct datasets, namely PIMA and Kaggle, through the utilization of various machine learning classification algorithms. The initial phase of the study involved crucial pre-processing steps, including data cleaning to address missing values. These steps ensured the integrity and reliability of the datasets.

The model selection process encompassed parameter and hyper parameter tuning using the Scikit Learn library within the Spider environment of Anaconda-3 2019 notebook. Among the different algorithms evaluated such as random forest, decision tree, SVM & ANN, but the artificial neural network (ANN) model stood out with an exceptional accuracy rate of 98.8%. This demonstrated a substantial improvement in accuracy and precision compared to existing diabetes prediction datasets. The ANN model showcased its potential to significantly enhance the accuracy and precision of diabetes prediction when applied to these specific datasets.

Furthermore, considering the success of the current research, there is potential for the application of this model in image processing to predict diabetes in both diabetic and nondiabetic individuals. By incorporating relevant image data, such as retinal scans or other medical imaging techniques, it may be possible to develop non-invasive and accurate methods for diabetes prediction. This avenue presents an exciting opportunity for future research and expansion of the current work, potentially revolutionizing the field of diabetes diagnosis and management.

Moving forward, it would be beneficial to explore the risk of diabetes development in nondiabetic individuals in the coming years. Additionally, there is potential to apply this model in image processing to predict diabetes in both diabetic and nondiabetic individuals. These avenues present exciting opportunities for future research and expansion of the current work.

Declarations

Acknowledgments:

I would like to express my sincere gratitude and appreciation to all those who have contributed to the successful completion of this research work. Firstly, I extend deepest appreciation to my research supervisor Mr. Yulin Wang, for their invaluable guidance, expertise, and continuous support throughout the project. Their insightful feedback and constructive criticism have greatly enriched the quality of my work.

Furthermore, I am grateful to the Wuhan University, Hubei China for providing me with the necessary resources, access to computing facilities, and a conducive research environment.

References

1. Jarrahi MH (2018) Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Bus Horiz 61(4):577–586
2. Dutta S (2018) An overview on the evolution and adoption of deep learning applications used in the industry. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1257
3. Fradella HF, Morrow WJ, Fischer RG, Ireland C (2010) Quantifying Katz: Empirically measuring reasonable expectations of privacy in the fourth amendment context. Am J Crim L 38:289
4. Yu X, Zhou S, Zou H, Wang Q, Liu C, Zang M, Liu T (2022) ”Survey of deep learning techniques for disease prediction based on omics data.” Hum Gene, 201140
5. Thakkar H, Shah V, Yagnik H, Shah M (2021) Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. Clin eHealth 4:12–23
6. Tigga N, Prerna, Garg S (2020) Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci 167:706–716
7. Viloria A, Herazo-Beltran Y, Cabrera D, Pineda OB (2020) Diabetes diagnostic prediction using vector support machines. Procedia Comput Sci 170:376–381
8. Sisodia D, Singh Sisodia D (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585
9. Swapna G, Vinayakumar R (2018) Soman.”Diabetes detection using deep learning algorithms. ICT express 4(4):243–246
10. Soman Kp (2018) Swapna, Goutham, and Ravi Vinayakumar. ”Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Comput Sci 132:1253–1262
11. Mujumdar A, Vaidehi V (2019) Diabetes prediction using machine learning algorithms. Procedia Comput Sci 165:292–299
12. Wu H, Yang S, Huang Z, He J, Wang X (2018) Type 2 diabetes mellitus prediction model based on data mining. Inf Med Unlocked 10:100–107

13. Chang V, Bailey J, Xu QA, Sun Z (2022) ” Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms.” Neural Comput Appl, 1–17
14. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) Predicting diabetes mellitus with machine learning techniques. Front Genet 9:515
15. El-Rashidy N, ElSayed NE, El-Ghamry A, Talaat FM (2022) Prediction of gestational diabetes based on explainable deep learning and fog computing. Soft Comput 26(21):11435–11450
16. Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, …, Abbas Z (2019) A model for early prediction of diabetes. Inf Med Unlocked 16:100204
17. Zhu C, Idemudia CU, Wenfang Feng (2019) Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inf Med Unlocked 17:100179
18. Khaleel FA, Al-Bakry AM (2023) ”Diagnosis of diabetes using machine learning algorithms.” Materials Today: Proceedings, 80, 3200–3203
19. Joshi TN, Chawan PPM (2018) ”Diabetes prediction using machine learning techniques ” Ijera 8(1):9–13
20. Kaur H, Kumari V (2022) Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inf 18(1/2):90–100
21. Khanam J, Jamal, Simon Y (2021) Foo. ”A comparison of machine learning algorithms for diabetes prediction. ICT Express 7(4):432–439
22. Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, Tiwari B (2022) ”A novel diabetes healthcare disease prediction framework using machine learning techniques”. Journal of Healthcare Engineering, 2022
23. Kamal CA, Atiyah MA (2023) ”Predict Diabetes Using Voting Classifier and Hyper Tuning Technique.” Kurdistan J Appl Res, 115–130
24. Rajput MR, Khedgikar SS (2022) Diabetes prediction and analysis using medical attributes: A Machine learning approach. J Xi’an Univ Archit Technol 14(1):98–103
25. Chou CY, Hsu DY, Chou CH (2023) Predicting the Onset of Diabetes with Machine Learning Methods. J Personalized Med 13(3):406
26. Alanazi A (2022) ”Using machine learning for healthcare challenges and opportunities.” Inf Med Unlocked, 100924
27. Nicolucci, A., Romeo, L., Bernardini, M., Vespasiani, M., Rossi, M. C., Petrelli,M. … Vespasiani, G. (2022). ”Prediction of complications of type 2 Diabetes: A Machine learning approach.” Diabetes Research and Clinical Practice, 190, 110013
28. Rastogi R, Bansal M (2023) Diabetes prediction model using data mining techniques. Measurement: Sens 25:100605
29. Malik MB, Ganie SM, Arif T (2022) Machine learning techniques in healthcare informatics: Showcasing prediction of type 2 diabetes mellitus disease using lifestyle data. Predictive Modeling in Biomedical Data Mining and Analysis. Academic Press, pp 295–311

30. Zelaya CV, Gonzalez (2019) ´”Towards explaining the effects of data preprocessing on machine learning.” 2019 IEEE 35th international conference on data engineering (ICDE). IEEE,
31. Ma N, Yu X, Yang T, Zhao Y, Li H (2022) ”A hypoglycemia early alarm method for patients with type 1 diabetes based on multidimensional sequential pattern mining”. Heliyon, 8(11), e11372
32. Han X, Chang L, Wang N, Kong W, Wang C (2023) Effects of Meteorological Factors on Apple Yield Based on Multilinear Regression Analysis”: A Case Study of Yantai Area. China Atmos 14(1):183
33. Edeh, M. O., Khalaf, O. I., Tavera, C. A., Tayeb, S., Ghouali, S., Abdulsahib, G.M. … Louni, A. (2022). A classification algorithm-based hybrid diabetes prediction model. *Frontiers in Public Health*, *10*, 829519
34. Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R (2023) Diabetes prediction using supervised machine learning. Procedia Comput Sci 216:21–30
35. Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi,L. … Deveci, M. (2023). ”A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion”. Information Fusion
36. Pisner DA (2020) and David M. Schnyer. ”Support vector machine.” Machine learning. Academic Press, pp 101–121
37. Alanazi HO (2017) Abdul Hanan Abdullah, and Kashif Naseer Qureshi. ”A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. J Med Syst 41:1–10
38. Pradhan N, Rani G, Dhaka VS, Poonia RC (2020) Diabetes prediction using artificial neural network. Deep Learning Techniques for Biomedical and Health Informatics. Academic Press, pp 327–339
39. Rahhal D, Alhamouri R, Albataineh I, Duwairi R (2022), June ”Detection and Classification of Diabetic Retinopathy Using Artificial Intelligence Algorithms”. In 2022 13th International Conference on Information and Communication Systems (ICICS) (pp. 15–21). IEEE
40. Siddiqui SA, Ahmad A, Fatima N (2023) IoT-based disease prediction using machine learning. Comput Electr Eng 108:108675

Figures

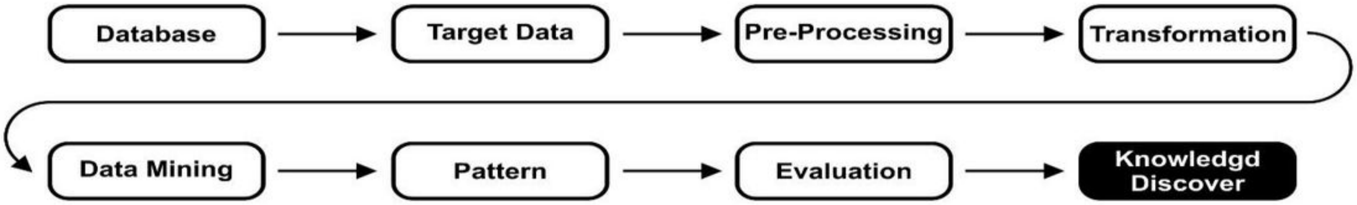

Figure 1

Process of Discover Knowledge

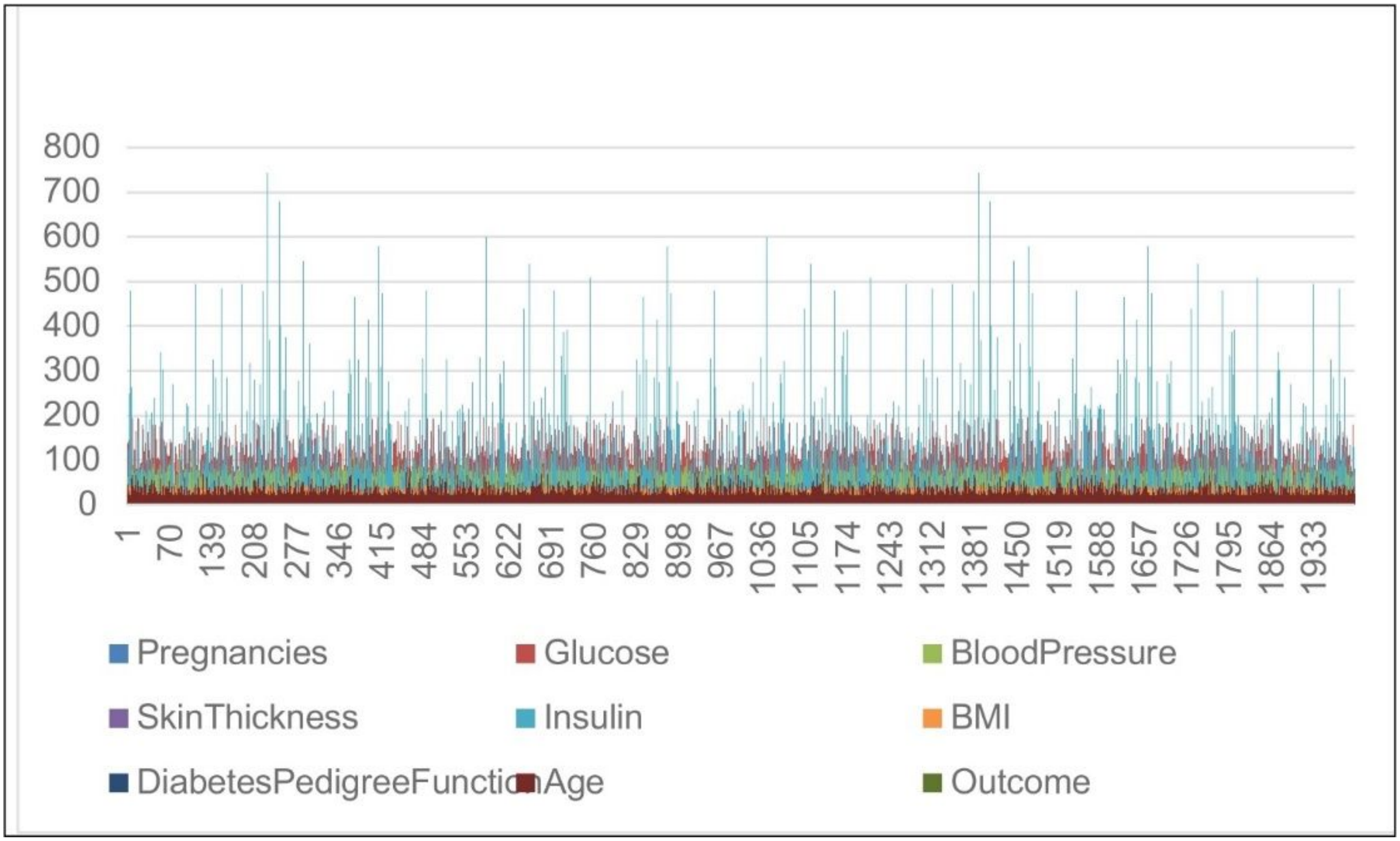

Figure 2

Graphical Analysis of Kaggle Diabetes Dataset

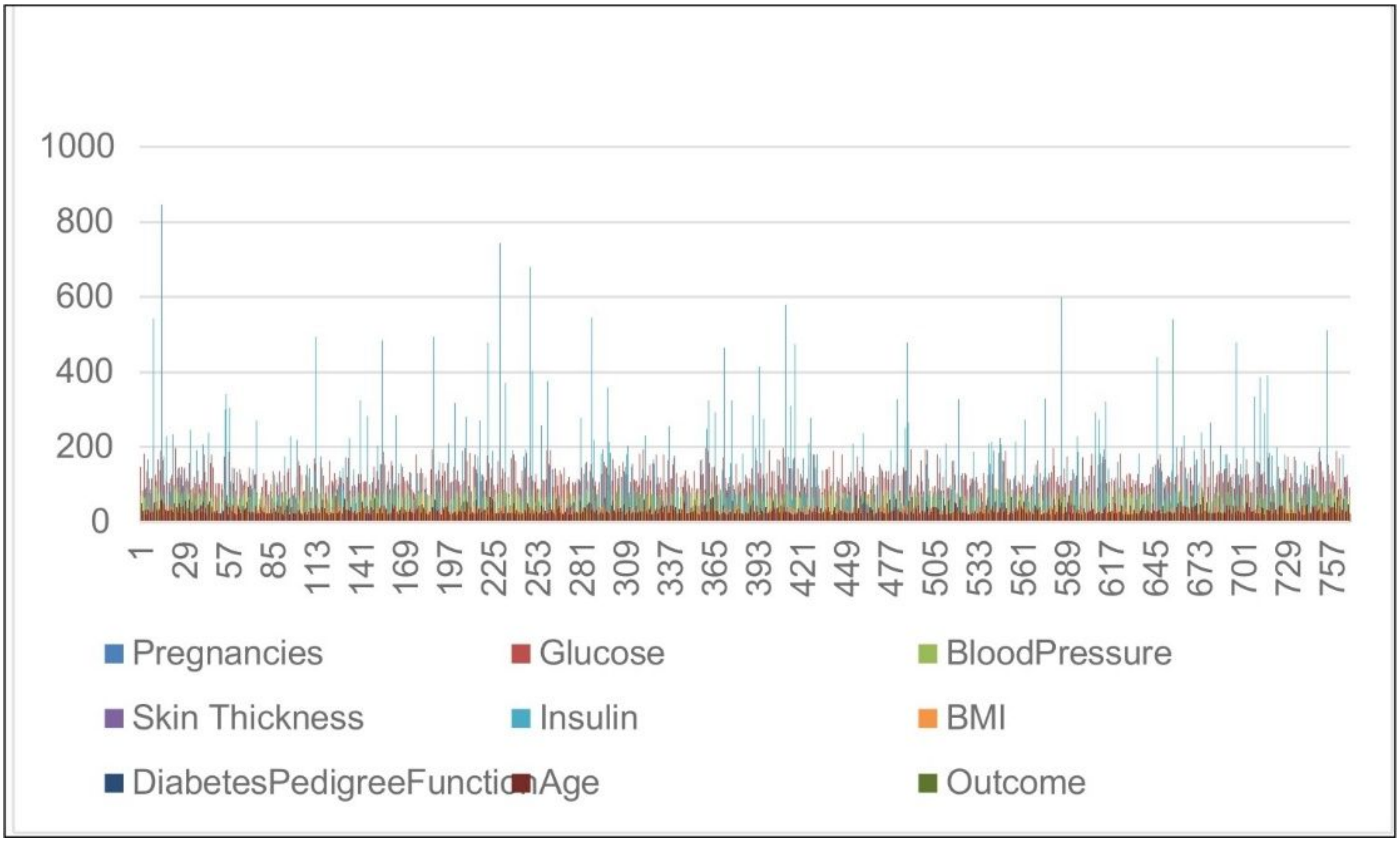

Figure 3

Graphical Analysis of PIMA Diabetes Dataset

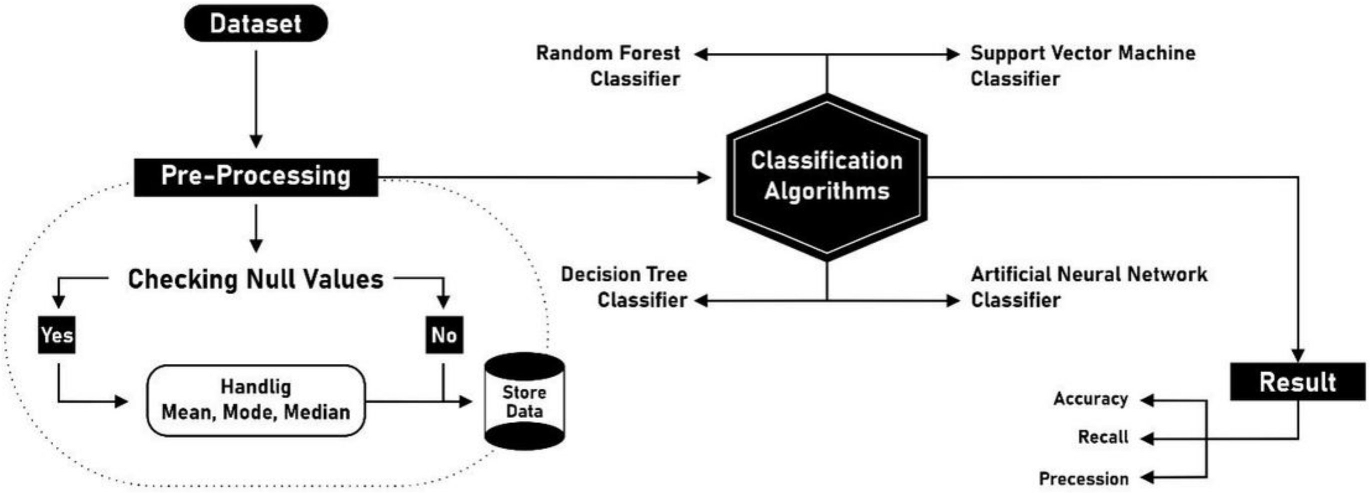

Figure 4

Proposed Model Architecture

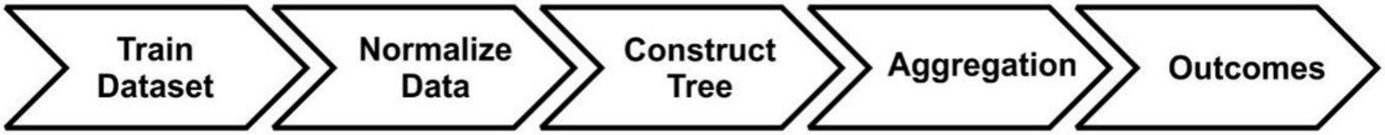

Figure 5

Sequential Manner of Random Forest

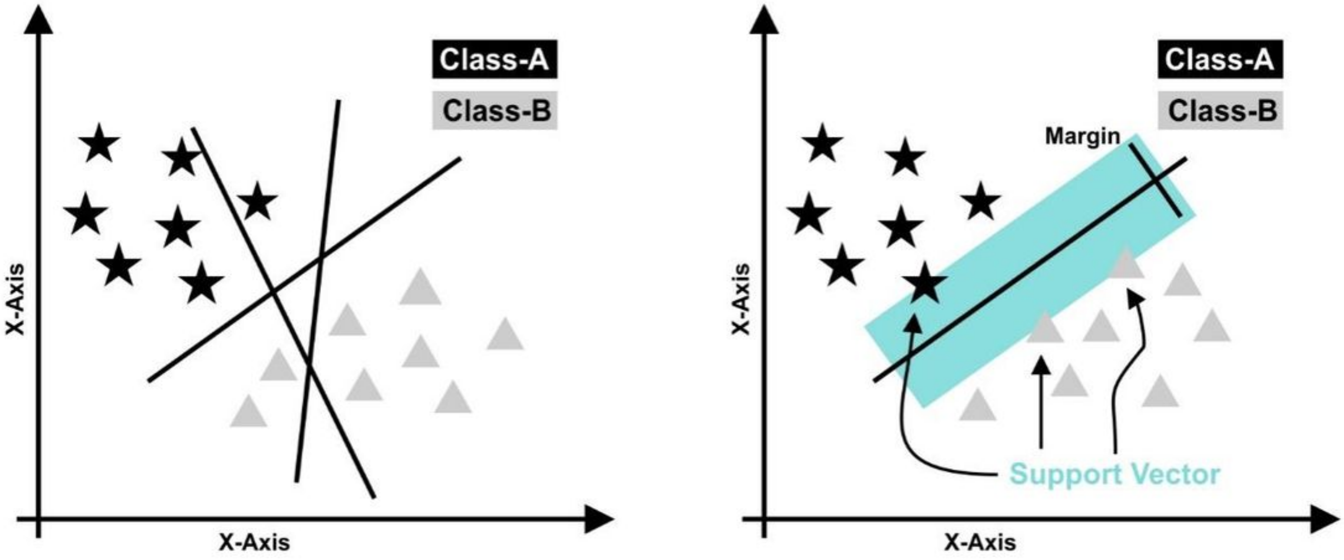

Figure 6

SVMs Hyperplane

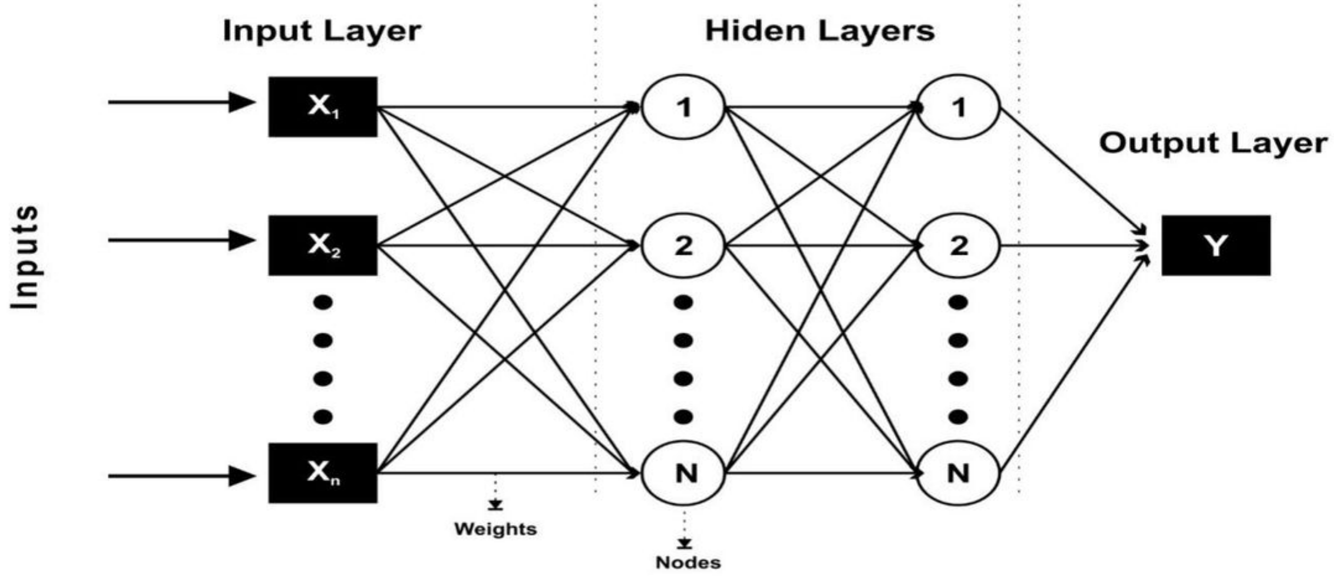

Figure 7

ANN Layer Architecture