[1]Sarita Kumari
[2]Dr. Amrita Upadhaya

Investigating Role of Supervised Machine Learning Approach in Classification of Diabetic Patient

Abstract*:* - Objectives: Healthcare analytics requires classifying diabetic patient datasets for quicker diagnosis and personalized treatment. This study used SVM, Decision Trees, KNN, ANN, and Logistic Regression to predict type 1 and 2 diabetes. Our detailed performance research shows these algorithms' utility in handling diabetic patient data's complexity.

Methods: We compare SVM, Decision Trees, KNN, ANN, and Logistic Regression for diabetes patient dataset classification. While each approach has pros and cons, ANN and logistic regression are promising clinical possibilities. Diagnoses, proactive therapies, and diabetes patient outcomes increase with these breakthroughs.

Results: SVM has 84.3% accuracy in type 1 and type 2 diabetics. SVM recognized complicated dataset patterns with great accuracy and recall. Decision trees were more interpretable and could record diverse choice limits, with accuracy rates of 86.15% for both types. With 90.6% accuracy, KNN predicted type 1 diabetes well. KNN was ideal for complex datasets because to its greater accuracy and recall using data point similarities. ANN and Logistic Regression had the highest accuracy for type 1 and type 2 diabetes patients at 96.1% and 97%, respectively.
Novelty: Layered ANN and logistic regression identified complicated dataset relationships with accuracy, recall, and F1 scores exceeding 0.94. ANN with logistic regression may change diabetes patient classification with unequaled prediction power and accuracy.

Keywords: Diabetic patient, Decision tree, SVM, Logistic Regression, ANN, KNN, Retinopath

I. Introduction

Recent developments in machine learning algorithms have been the driving force behind extraordinary progress in the field of healthcare, namely in the categorization of datasets including diabetic patient information. A number of different types of (ANN), Decision Trees (KNN), (SVM) and Logistic Regression are considered to be the cornerstones of innovation in this field. Because of its ability to do non-linear classification and its resistance to overfitting, (SVM) have shown their effectiveness in identifying detailed patterns within the data of diabetes patients. In a similar vein, Decision Trees provide interpretability and adaptability, which enables physicians to extract useful insights from complicated information. In the meanwhile, the straightforwardness and intuitiveness of the KNN approach have made it possible to accurately classify patients by capitalizing on the commonalities that exist between the data points. Furthermore, the introduction of artificial neural networks (ANN), which has the capacity to understand complicated associations in data via layered architectures, has revolutionized the categorization of diabetes patients, resulting in an accuracy and prediction capability that has never been seen before. Even though these significant gains have been made, there are still identifiable gaps that need additional exploration. A number of important directions for investigation include the requirement for greater interpretability in complicated models, the difficulty of managing unbalanced datasets that are inherent in medical diagnostics, and the pursuit of enhanced generalization skills over a wide range of patient groups. Not only can addressing these gaps improve the effectiveness of classification models, but it also helps to nurture the real-world usability of these models in clinical situations. Therefore, the purpose of this research is to strive to bridge these gaps by providing a nuanced

[1] *Corresponding author: Banasthali Vidiya Peeth, Jaipur Rajasthan (INDIA)
[2] Banasthali Vidiya Peeth, Jaipur Rajasthan (INDIA)

view on the categorization of diabetes patients while simultaneously extending the frontiers of ML in the healthcare industry.

Diabetes

People with diabetes will always have trouble converting the food they consume into energy. Glucose is the primary form of sugar absorbed into your bloodstream from eating [1]. When blood glucose levels rise, the pancreas secretes insulin [2]. One of the most important hormones, insulin, lets glucose into cells so that they may be used as fuel. One of the two main complications of diabetes is an inability to properly use the insulin that the body does generate [3]. When cells either cease reacting to insulin or insulin production is inadequate, the result is an accumulation of glucose in the blood. Chronic diseases might progress to more serious conditions, including heart disease and renal failure [4]. It should be noted that several types of diabetes exist [5]. Typically, you may say that they fall into one of these categories:

- *Type 1 Diabetes:* Type 1 diabetes is believed to be caused by an immunological response. Insulin secretion is reduced as a result of this reaction. Type 1 diabetes affects only around 5-10% of people with diabetes. The onset of symptoms is often rather quick in those diagnosed with type 1 diabetes. The usual patient is a young person, often in their teens or early twenties [6]. Injecting oneself with insulin every day is a must for those living with type 1 diabetes. To this day, there is still no cure for type 1 diabetes.

- *Type II Diabetes:* Regular insulin treatment does not alleviate the persistently high blood sugar levels seen by people with type 2 diabetes. Type 2 diabetes accounts for around 90% to 95% of all diabetes cases. Because it takes years to manifest, diagnosis is often reserved for adults. They should be required to monitor blood sugar levels regardless of whether they are unwell or not if we are in danger [7]. Modifying one's eating choices, increasing physical activity, and avoiding or delaying the start of type 2 diabetes are all possible outcomes of adopting a healthier lifestyle.

- *Gestational (Type 3) diabetes:* Hyperglycemia, or gestational diabetes, is a condition when the mother's blood glucose levels are increased but not high enough to be diagnosed as diabetes. During pregnancy, gestational diabetes often becomes apparent. Consequently, complications during pregnancy and delivery are more common in women who have gestational diabetes [8]. There is an increased risk of type 2 diabetes in these mothers and maybe their offspring. Instead than relying on patient self-reports, the most reliable way to detect gestational diabetes is to do prenatal screenings for risk factors. We trained the machine-learning model on this mountain of medical data, then drew some insights using data visualization and analysis.

II. Machine Learning

With the help of ML, computers can learn to make decisions on their own. These decisions are made by the computer when it understands the patterns in the data and learns from them [9]. The outcome, in the form of a prediction or a classification, is then produced by means of pattern matching along with further analysis. Types of ML include:

- *Supervised learning:* In the realm of machine learning, supervised learning reigns supreme. When the input-output data is perfectly mapped, it is often used. A ML algorithm is the approach that the AI system uses to do its goal, which is often to make predictions based on input data [10]. Machine learning algorithms mainly use two methods: classification and regression.
- *Unsupervised Learning:* The machine is given input data and is then allowed to derive its own conclusions in unsupervised learning. Due to the absence of classifications in the unlabeled dataset, unstructured models may be trained without human intervention from the programmer [11]. Without human oversight, a model may sift through mountains of data in search of patterns and insights in unsupervised learning. Their application is in the resolution of Association and Clustering issues.
- *Reinforcement Learning:* Entities participate in reinforcement learning when they take in data from their actions and change their behavior based on that data. Incentives, such as praise for good conduct and reprimand for bad, serve as reinforcement for the entity's actions [12]. There is zero oversight of the representative. Reinforcement learning makes use of the Q-learning technique.

A. SVM (Support Vector Machine)

One well-liked directed learning technology, (SVMs) have dual use: they can classify data and solve regression problems. On the other hand, classification tasks are where it is most often used in the field of ML [13]. The (SVM) method is developed to provide best boundary or line that can partition n-dimensional space into groups, allowing for future accurate categorization of fresh data points. The most significant decisions may be taken within the confines of a hyper-plane [14]. This is why the SVM may look like this:

- *Linear SVM:* Linear (SVMs) are a kind of classifier that may be used when a dataset can be easily partitioned into two sections by drawing a straight line.
- *Non-linear SVM:* We utilize a classifier known as a Non-linear SVM to find datasets that don't neatly fit into a linear hierarchy. For regression and classification jobs, supervised machine learning algorithms like (SVM) come in handy. Its categorization issues are its specialty. Finding the optimal hyper plane to maximize the margin between data points of distinct classes is main objective of SVM. SVM operates as follows:
- *Data Preparation:* An example of a labeled dataset would be one in which each data point has a class label (a positive number for continuous classification or a negative number for binary classification). Multiclass issues may also be addressed by extending SVM. If your data isn't up to par, you may fix it by doing feature selection or extraction.
- *Model Training:* In (SVMs), the goal is to maximize margin, which is distance between hyper plane along with closest data points of each class, by finding the hyper plane that best separates data points. An optimization issue is solved to do this. The primary goal is to locate data points that are near the decision border, known as support vectors.
- *Classification:* The algorithm determines the data point's classification based on its location relative to the hyper plane. It is allocated to one class if it is on one side and to the other class

if it is on the other side.(SVMs) have found extensive use in several domains, such as bioinformatics, image and text categorization, and many more. Nevertheless, their performance may be greatly affected by the kernel and hyper parameters you choose, so it's important to tune them carefully for best results.

B. Decision tree

One common model for data categorization is the decision tree, which resembles a tree. Dividing data into branches, which may be other trees or even just plain old nodes, is what a decision tree does [15]. There are three possible types of nodes in a decision tree. Here, we see that the node contains several branches. Deciduous nodes at the tree's base are called leaf nodes.

- *Root Node:* This is still another decision node at the very top level [16]. When it comes to classification and regression, a supervised machine learning approach that is often used is a decision tree. It is a flexible and interpretable method that selects subsets of the dataset iteratively according to the most important feature at each node. Decision tree algorithm is shown here:
- *Data Preparation:* Each data point in a labeled dataset contains a collection of characteristics and a corresponding target label. This is the first step.
- *Tree Construction:* To start, the algorithm takes into consideration all of the characteristics and chooses the one that gives the best split. It checks a variety of split criteria, the most popular of which are Gini Impurity and Mean Squared Error. The feature that was chosen serves as the tree's decision node.
- *Splitting:* Based on the values of the feature that was chosen, the dataset is separated into subsets. In a tree structure, each subset stands in for a different branch or child node.
- *Recursive Splitting:* When a stopping criterion is reached, such as a maximum tree depth, a minimum digit of samples in a node, or no further progress in impurity or error reduction, Steps 2 and 3 are repeated for each child node.
- *Leaf Nodes:* The last nodes of a tree, known as leaf nodes or terminal nodes, are located when a stopping requirement is achieved. These nodes stand for the expected value or class.
- For example, decision support systems, categorization, and regression are just a few of the many common uses for decision trees. Because of their interpretability and simplicity, they are often used as a foundation for more intricate algorithms.

C. K-Nearest Neighbors

Classification problems often use the nonparametric sluggish supervised learning technique NN. Although the K-NN is complex, the two most salient aspects are as follows: Just so we're clear, KNN isn't making any assumptions on the data distribution since it's a nonparametric method. Since it is a lazy learner approach, KNN does not teach algorithm to differentiate between classes using the training dataset [17]. To sidestep this problem, a lazy learner algorithm commits training dataset to memory along with refrains from abstracting the data until predicted results are needed. An instance-based machine learning technique for classification and regression, (K-NN) is straightforward and non-parametric. To generate predictions, K-NN locates the "k" data points in training dataset that are geographically closest to a certain test data point. In the case of classification, the output is determined by a majority vote, and in the case of regression, it is determined by an ever-aging technique. This is how K-NN operates:

For Classification:

- *Data Preparation:* Get started using a labeled dataset that contains features and class labels attached to each data point. Choose a Value for K: Choose value of "k," number of neighbors to be taken into account for prediction purposes. This value is usually an odd integer so that there are no ties.
- *Distance Metric:* To determine how similar two sets of data are, choose a distance metric (e.g., Manhattan distance, or the Euclidean distance).
- *Prediction:* In order to assign a label to a newly-added data point, one must first determine its distance from every other data point in the training set. Pick the "k" smallest distance data points. Use a majority vote among the k-nearest neighbors to decide the class label.

For Regression:

- *Data Preparation:* The first step is to create a dataset containing features and a numerical goal value assigned to each data point. Set K to a Value: Choose the value of "k," the number of neighbors to be taken into account for prediction purposes. This value is usually an odd number, just like in categorization.
- *Distance Metric:* Pick a distance metric to see how close the data points are to each other.
- *Prediction:* Finding distance between a new data point along with every other data point in training dataset allows you to make a numerical prediction about that point.

D. Artificial Neural Network

To describe a system that can learn and adapt to new environments in a way similar to how the brain develops in the embryonic stage, the phrase "Artificial Neural Network" was developed [18]. In (ANN), the interconnections between layers of "neurons" mimic those between actual brain cells. Nodes are the proper names for these clusters of neurons. The architecture and operation of biological neural networks, like the brain, serve as inspiration for ANNs, a category of machine learning algorithms. Classification, regression, pattern recognition, and many more applications using ANNs. They are an essential part of the machine learning branch known as deep learning. A brief explanation of how artificial neural networks function is as follows:

Neurons and Layers: Synthetic neurons, or ANNs, are networks of linked nodes. The features or data are fed into the input layer, and the predictions or results are output into the output layer.

Connections and Weights: Connected neurons in one layer communicate with neighboring neurons in the next layer. The strength of a link is determined by its related weight. During training, the weights are adjusted to maximize the network's performance; they are initially given random values.

- *Activation Function:* An activation function processes weighted total of each neuron's inputs. By making the network non-linear, activation function enables it to represent intricate data associations. Some common activation functions include the sigmoid, the (ReLU), and (tanh).
- *Feed forward Propagation:* The input data is transferred layer by layer across the network during the forward pass, also known as feed forward propagation. After applying the activation function and weighting the inputs, neurons in each layer calculate their outputs. The prediction made by the network is output by the last layer.
- Loss Function: The loss function is a metric for gauging how far off the goal values are from the predictions made by the network. Depending on the job at hand, a specific loss function might be used. In gradient descent and back propagation, the network learns by lowering the loss function via weight adjustments. Gradient descent and other optimization algorithms are

often used for this. In order to update the weights repeatedly, one uses back propagation to determine the loss gradients with regard to the weights.

- Training: Training entails feeding the network the training data, making predictions, determining the loss, and then changing the weights via back propagation. Once the loss reaches a certain threshold, or after a certain number of epochs, the training process terminates.
- *Hyper parameter Tuning:* A variety of hyper-parameters of ANNs, such as learning rate, the digit of layers, the digit of neurons in each layer, and the choice of activation functions, must be fine-tuned. The network's performance may be optimized using hyper parameter adjustment. Regularization methods like as batch normalization, L1 and L2 regularization, and dropout are often used to avoid over-fitting.
- *Evaluation and Testing:* A different test dataset is used to evaluate the trained network's generalization capability. Performance is evaluated using metrics that are pertinent to the work at hand, such as accuracy and mean squared error.
- Numerous fields have seen the astounding success of Artificial Neural Networks, such as driverless cars, natural language processing, picture and audio recognition, and more. Recent years have seen tremendous progress in machine learning thanks to deep learning, which entails training deep neural networks with several hidden layers.

E. Logistic Regression

For binary classification problems, logistic regression is a statistical approach is used. In these types of tasks, the outcome variable is categorical and has only two potential outcomes, which are often denoted by the numbers 0 and 1. On the basis of more predictor variables, it is a form of regression analysis that is often used in the field of ML for purpose of predicting the chance of experiencing a binary result. The following is a condensed explanation of how the

- *logistic regression method operates:* The output of a linear equation is transformed into a number that falls somewhere between 0 and 1 by the use of the sigmoid function, which is also referred to as the logistic function. Logistic regression makes use of this function.
- *Decision Boundary:* Logistic regression typically makes a prediction of class 1 if estimated probability is higher than or equal to 0.5, and class 0 if estimated probability is less than or equal to 0.5.
- *Training:* The logistic regression model is trained by calculating the coefficients that provide the greatest fit for the training data. In many cases, this is accomplished via the use of optimization strategies such as maximum likelihood estimation or gradient descent. Due to the fact, it is easy to understand and straightforward to implement, it is considered a basic approach in the area of machine learning and statistics.

III. Related Work

This literature review compiles the results of healthcare dataset analyses and predictions made using various methods. Many different prediction models, sometimes combining data mining and machine learning techniques, have been developed and used by researchers. Using a well-established and trustworthy ML voting method, Mahabub A. (2019) investigated possible diabetic complications. Presenting a comprehensive list of ML along with DM usage in context of diabetes was the objective of this paper [1]. Land Cover Classification Sentinel-2 Satellite Imagery Using

Support Vector Machines along with Radial Basis Functions was the main emphasis of Thanh Noi P's (2017) research. Accuracy was high for both balanced and skewed datasets [2]. Ensemble techniques for type 2 diabetes were presented by Alehegn M (2019). The 130 US hospital diabetes data sets and PIDD are used in this study's approach [3]. DT, KNN, along with RF Classifier as Models in case of Predicting Bookings for FSBO (Afrianto MA, 2020). Area under curve for all models studied AUC-ROC was greatest for Random Forest Classifiers, according to the results [4]. The naive-Bayes classifier was studied by Agnal AS (2020) in relation to diabetes data evaluation. It is clear from the graph that diabetes has been on the rise in both incidence and prevalence over the last several decades [5]. According to Alghurair NI's (2020) assessment, generic frameworks might be useful in case of SVM, ANN, LGBM, along with LR. After comparing the provided frameworks to other state-of -art options, the second one came out on top [6]. In Saudi Arabia, Almutairi ES (2023) examines the effectiveness of several categorization methods for diabetes prevalence rates along with predicted changes in the disease in relation to pertinent behavioral risk factors (smoking, obesity, along with inactivity). Specifically, the study focuses on the prevalence rates of diabetes in Saudi Arabia [7]. Bansal M, (2022) took into consideration the majority of the features pertaining to five machine learning algorithms, namely KNN, GA, SVM, DT, along with LSTM network. These algorithms have been covered in great depth, which is a necessity for entering the area of machine learning [8]. The technique that Charbuty B, (2021) takes to the decision trees is both comprehensive and thorough. Moreover, the particulars of the study, including the algorithms and techniques that were used, the datasets that were utilized, and the results that were reached [9]. The research conducted by Kiranashree BK, (2021) concentrated on several machine learning algorithms and physiological factors for purpose of stress detection [10]. Kumari S. (2024) was given responsibility of comparing and contrasting overall of SVM, Decision Tree, KNN, along with ANN on a dataset that included diabetes patient classifications. Acquisition of more accurate along with trustworthy results is one aim of the study [11]. Researchers D. F. M. Mohideen et al. (2021) used a combination of regression imputation and a Gaussian Naive Bayes algorithm to accurately forecast occurrence of diabetes mellitus [12].

IV. Research Gap

Despite the fact that a multitude of studies have investigated the use of various ML algorithms, such as (SVM), (KNN), (ANN), along with Logistic Regression in classification of diabetic patient datasets, there is still a significant gap in research literature concerning a comprehensive comparative analysis of these algorithms that is specifically tailored to the field of diabetes prediction. Furthermore, the bulk of research focus solely on basic machine learning algorithms, ignoring the possibility of innovations or alterations within these algorithms that might have a considerable influence on the prediction accuracy, interpretability, and generalization capabilities of the proposed system. For example, while some research investigates ways to increase the performance of standard support vector machine (SVM) models by developing enhanced non-linear kernels, other research investigates innovative feature selection strategies or hybrid optimization approaches to improve classification performance. Furthermore, there is still a lack of exploration about the process of translating research results into clinical practice. In spite of the encouraging findings that have been published in the academic literature, there is a scarcity of research that investigates the applicability and scalability of ML models for diabetes prediction in the real world, taking into account a wide range of healthcare settings, patient groups, and data gathering techniques. To make progress in the area of diabetes patient categorization, to facilitate informed decision-making by healthcare practitioners, and eventually to improve patient outcomes via individualized medical treatments, it is essential to address the research gaps that have been identified. By performing a complete comparative analysis of SVM, Decision Tree, KNN, ANN along with logistic regression algorithms in classification of diabetic patient datasets, as well as evaluating their practical value and scalability in clinical settings, the purpose of this work is to bridge the gaps that have been identified.

- *Problem Statement:* The topic of diabetic patient data classification has been the subject of several research. But traditional studies have trouble with time and space consumption because of visuals that are too big. Using SVM, we were able to predict a DM diagnosis based on patient-reported variables. In terms of the outcome variable, there are three potential states: none, at risk in case of developing diabetes, along with diagnosed diabetes. Predicting instances of diabetes involves the use of classification algorithms. Various approaches in case of diabetes prediction with many hidden layers are used and evaluated in study.
- *Dataset Used:* Our focused in this review article is on methods for data collection and analysis that might help with future forecasting and product assessment by revealing trends and patterns. What follows is a synopsis of the dataset. We have to provide a more thorough analysis based on the publications included in Table 1, which represent previous work in this area and serve as the basis for this review study.

Table 1. Dataset Information

Name	Link	Description	Format
Diabetes Patients Data	https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset	Research supported by the (NIH) is the primary source for the majority of this data.	From the data set in (.csv) File, along with can find several variables.
Diabetes UCI Dataset	https://www.kaggle.com/datasets/alakaaay/diabetes-uci-dataset	Research supported by National Institute of Diabetes along with Digestive along with kidney diseases provides the bulk of this data. Confirmation of diabetes in a patient is the goal of these examinations.	From data set in (.csv) File.
Diabetes Health Indicators Dataset	https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset	The BRFSS is a yearly telephone survey that pertains to health and is collected by the CDC.	Diabetes _ 012 _ health _ indicators_ BRFSS2015.csv is a clean dataset of 253,680 survey responses.

Diabetes Disease Updated Dataset	https://www.kaggle.com/datasets/jillanisofttech/diabetes-disease-updated-dataset	Research supported by National Institute of Diabetes along with Digestive along with Kidney Diseases provides the bulk of this data.	From data set in (.csv) File.
Diabetes Dataset for Beginners	https://www.kaggle.com/code/melikedilekci/diabetes-dataset-for-beginners	Every dataset consists of a single outcome variable and many medical prognostic variables.	Dataset in (.csv) File.

V. Significance Of Research

There are significant benefits that this study brings to the field of healthcare analytics, specifically in the area of diabetes patient categorization. Provides valuable insights into performance of four prominent ML algorithms, namely (SVM), Decision Trees, KNN, ANN, along with Logistic Regression in predicting cases of type 1 along with type 2 diabetes. This was accomplished by conducting a comprehensive comparative analysis of these algorithms. This empirical study not only offers healthcare practitioners ideas that can be put into action for the selection of algorithms and the building of models, but it also highlights the potential of ML approaches in improving diagnostic accuracy and patient care in the context of diabetes treatment. In addition, the study has immediate consequences for clinical practice, since it lays the framework for the creation of decision support systems and prediction models that will assist doctors in the early identification of diabetes patients and in the tailored treatment planning of diabetic patients. Furthermore, the study adds to the advancement of knowledge and understanding in healthcare analytics by defining ideal algorithms and detailing outlook research areas.

VI. Methodology

In the section, we covered research approach. Several traditional methods have focused on data categorization. The current study aimed to classify datasets of diabetes patients by investigating the function of SVM, decision trees, KNN, ANN, along with logistic regression. Conventional classification strategies have been investigated in the literature. A dataset of diabetic patients has been used to simulate accuracy using SVM, decision trees, KNN, ANN, and logistic regression in order to assess dependability of classifiers. The major research process flow is shown in Figure 1.

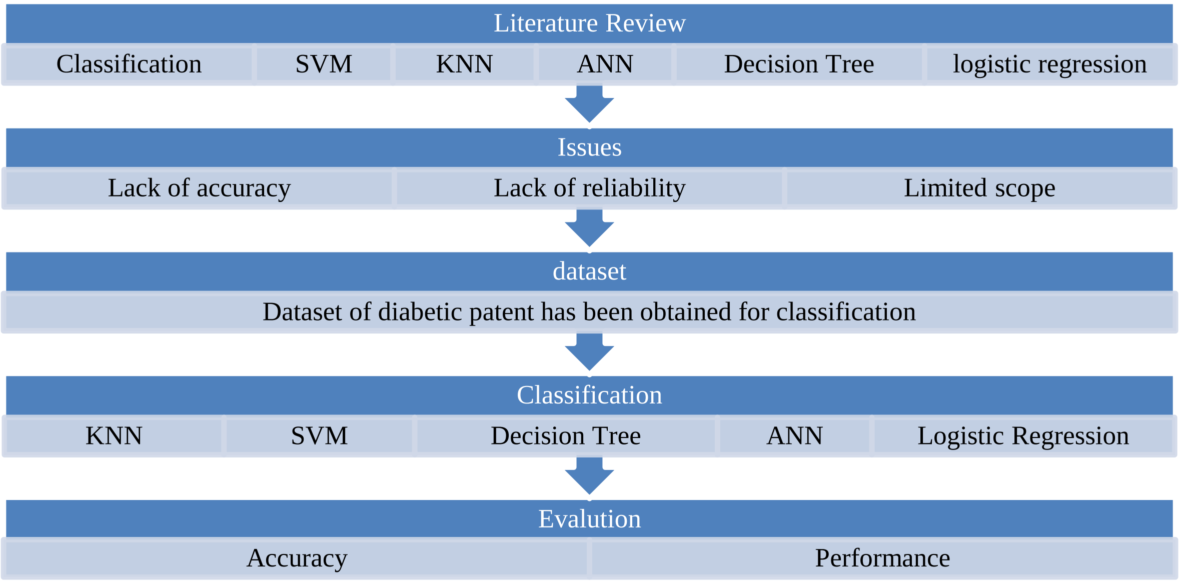


Figure 1: Flow of work

When compared to more traditional classification methods like SVM, Decision tree, and KNN, research has shown that ANN and logistic regression provide superior accuracy. The training dataset in the ANN and logistic regression model, as well as the process flow for obtaining accuracy, are shown in Figure 2. Everything we did was vital in achieving our goals and answering our questions.

Comparison of Accuracy of Conventional ML Approaches:

A matrix is used to summarize and display the model's overall success. Here are the outcomes we achieved by applying several ML algorithms on the dataset. The most accurate technique is Logistic Regression, which has a precision of 96%. The accuracy for several classifiers is shown in Table 2.

Table 2. Accuracy Tab

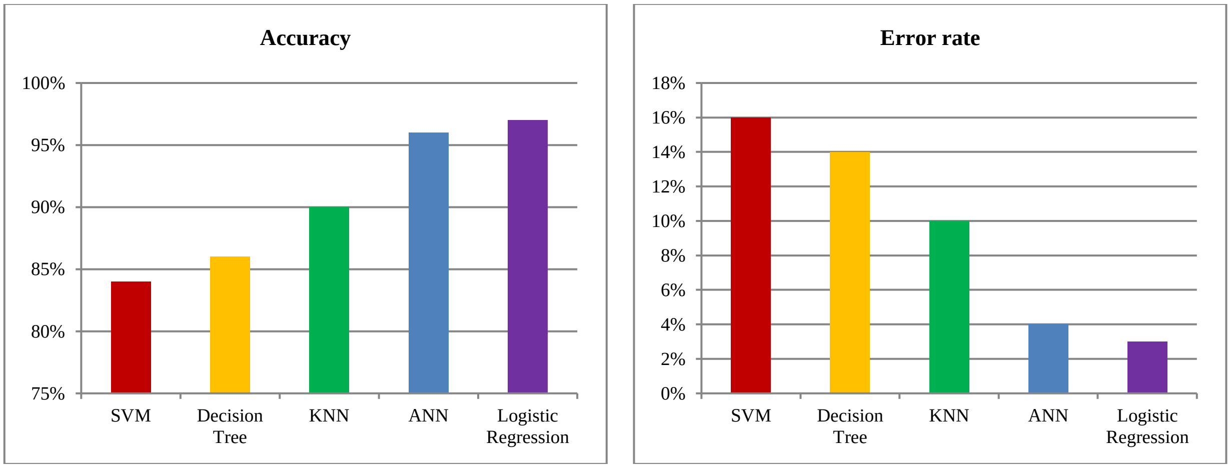
Algorithms	Accuracy
Decision Tree [6]	86%
Gaussian NB [12]	93%
LDA [18]	94%
SVC [18]	60%
Random Forest [1]	91%
Extra Trees [13]	91%
AdaBoost [18]	93%
Perceptron [18]	76%
Logistic Regression [5]	96%
Gradient Boost Classifier [18]	93%
Bagging [18]	90%
KNN [2]	90%

In Table 3 presents outcome of simulation in case of 4 mechanisms that are used for classification of diabetic dataset. As shown in table below:

Table 3. Comparison of Accuracy and Error rate in case of Different Classification Mechanisms

Mechanism	Accuracy	Error rate
SVM	84%	16%
Decision Tree	86%	14%
KNN	90%	10%
ANN	96%	4%
Logistic Regression	97%	3%

Table 3 shows that numerous classifiers are compared using the same datasets, which are diabetes datasets. In that comparison, the accuracy scores of SVM, Decision Tree, KNN, ANN, and Logistic Regression were 84%, 86%, 90%, 96%, and 97%, respectively. When compared to other classifiers, ANN produces superior results in terms of their accuracy. Once the optimal accuracy has been determined, Table 3 displays a comparison of the categorization technique's error rates. Table 4 shows that numerous classifiers are compared using the same datasets, which are diabetes datasets. The error rates for SVM, Decision Tree, KNN, ANN, and Logistic Regression are 16%, 14%, 10%, 4%, and 3%, respectively, when compared. The lowest error rate is 4% for ANN and the worst is 16% for SVM.



(a) (b)

Figure 2. Comparison of (a) Accuracy and (b) Error Rate for different classification mechanisms

Accuracy comparisons of several categorization techniques are shown in Figure 2 (a). All of the classifier's accuracy, properly and accurately, is summarized in the provided chart. To ensure accurate graph representation, include Table 3 into entries. The study produces a matching chart in Figure 2 (b) and displays error rate of each classifier in tabular form. To ensure accurate graph representation, include Table 4 into entries. To Predict Type1 and Type2 Diabetic Predicting whether an individual has Type 1 or Type 2 diabetes is a classification task can be approached using various ML algorithms, including ANNs and logistic regression. To create a predictive model for this task, you would need a dataset containing features. For each patient, as well as labels indicating whether they have Type 1 or Type 2 diabetes. Here are the steps to build a predictive model using ANNs for this task:

- *Data Collection:* Collect a dataset with relevant features and labels. The dataset should include patient characteristics.
- *Data Preprocessing:* Take care of outliers along with missing numbers to clean up the data. In case it's needed, encode category variables. To make all numerical characteristics use the same scale, normalize or standardize them.
- *Data Splitting:* Split dataset into a training set along with a testing set. This allows you to train and evaluate model's performance on different data.
- *Model Architecture:* Design architecture of your ANN. For binary classification like this, a simple feed forward neural network with one or more hidden layers can be a good starting point. Decide digit of neurons in each layer along with choose appropriate activation functions.
- *Compile the Model:* Choose a loss function suitable for binary classification. Select an optimizer. Define evaluation metrics for monitoring the model's performance during training.
- *Training:* Train the ANN and logistic regression on the training dataset using back propagation and gradient descent. Monitor training process and evaluate model's performance on the validation set to detect over fitting. Adjust hyper parameters and add regularization techniques as needed.
- *Evaluation:* After training, evaluate model's performance on test dataset to assess its ability to generalize to new, unseen data. Calculate various evaluation metrics to measure model's accuracy, precision, recall, F1-score, along with ROC curve if applicable.
- *Tuning and Optimization:* Fine-tune model by adjusting hyper parameters along with exploring different network architectures to achieve better performance.
- *Deployment:* Once satisfied with models' performance, you can deploy it in a clinical setting, provided it meets the necessary regulatory and ethical requirements.
- *Monitoring and Maintenance:* Continuously monitor model's performance along with retrain it periodically with updated data to maintain its accuracy. Remember that medical applications, especially those related to diagnosis and treatment, require rigorous validation and ethical considerations. It's essential to collaborate with healthcare professionals, follow relevant regulations, and ensure the model's safety and reliability before deploying it in a real-world medical environment.

VII. Result And Discussion

Results must be validated with previous results on relevant literatures and to be discussed with favoring and contradicting the current findings. Comparison with existing reports should be provided in order to prove validity of present model or program developed

a. Confusion Matrix for different Machine learning

To effectively validate the results obtained from the classification models utilizing SVM, Decision Tree, KNN, and ANN algorithms for predicting type 1 and type 2 diabetic patients, a thorough comparison with existing literature is essential. Table 4 displays the confusion matrices for (SVM), Decision Trees, (KNN), (ANN) and logistic regression algorithms when used to predict instances of type 1 along with type 2 diabetes. SVM matrix indicates that out of a total of 2000 samples, 799 instances of type 1 diabetes were accurately diagnosed, whereas 113 were mistakenly categorized as type 2. In the same manner, a total of 887 individuals with type 2 diabetes were accurately recognized, whereas 201 instances were incorrectly categorized as type 1 diabetes. The Decision Trees demonstrated similar performance by properly categorizing 842 instances of type 1 diabetes and 881 cases of type 2 diabetes. Nevertheless, it erroneously categorized 119 instances as type 1 cases and 158 instances as type 2 cases. The K-Nearest Neighbors (KNN) algorithm is shown strong and consistent performance, especially in accurately detecting instances of type 1 diabetes. Out of a total of 930 cases, 871 were properly recognized while just 59 were misclassified. In the case of type 2 diabetes, 941 patients were accurately categorized, whereas 129 instances were classified incorrectly. On the other hand, Artificial Neural Network (ANN) has shown outstanding performance by precisely categorizing 941 instances of type 1 diabetes and 981 cases of type 2 diabetes. A total of 19 instances of type 1 were misclassified, along with 59 cases of type 2. And, logistic regression shown outstanding performance by precisely categorizing 941 instances of type 1 diabetes and 981 cases of type 2 diabetes. A total of 19 instances of type 1 were misclassified, along with 59 cases of type 2.

Table 4. Confusion matrix of SVM, Decision Tree, KNN along with ANN algorithms to predict type 1 along with type 2 diabetic

		Type 1	**Type 2**
SVM	Type 1	799	113
	Type 2	201	887
Decision Tree	Type 1	842	119
	Type 2	158	881
KNN	Type 1	871	59
	Type 2	129	941
ANN	Type 1	941	19
	Type 2	59	981
Logistic Regression	Type 1	958	14
	Type 2	42	986

In general, the confusion matrices demonstrate the different levels of performance shown by the four methods. Although SVM and Decision Trees produced fairly even outcomes, KNN revealed exceptional proficiency in categorizing type 1 instances, while ANN and Logistic Regression exhibited better performance in both kinds. These findings emphasize the significance of choosing the suitable algorithm that aligns with the unique criteria and features of the dataset in order to get highest level of accuracy in predicting categorization of diabetes patients.

b. Accuracy parameter for different Machine learning

Confusion matrices depicting the performance of each algorithm in predicting type 1 and type 2 diabetic cases can be presented side by side, allowing for a visual comparison of their classification accuracy, precision, recall, along with F1-score. Additionally, statistical measures such as accuracy, along with sensitivity, specificity, along with area under the ROC curve AUC can be calculated along with compared across different algorithms to provide a comprehensive evaluation of their predictive performance. Table 5 presents a summary of the accuracy parameters for (SVM), Decision Trees, (KNN), (ANN) and Logistic Regression techniques.

Table 5. Accuracy parameter of SVM, Decision Tree, KNN along with ANN algorithms to predict type 1 and type 2 diabetic

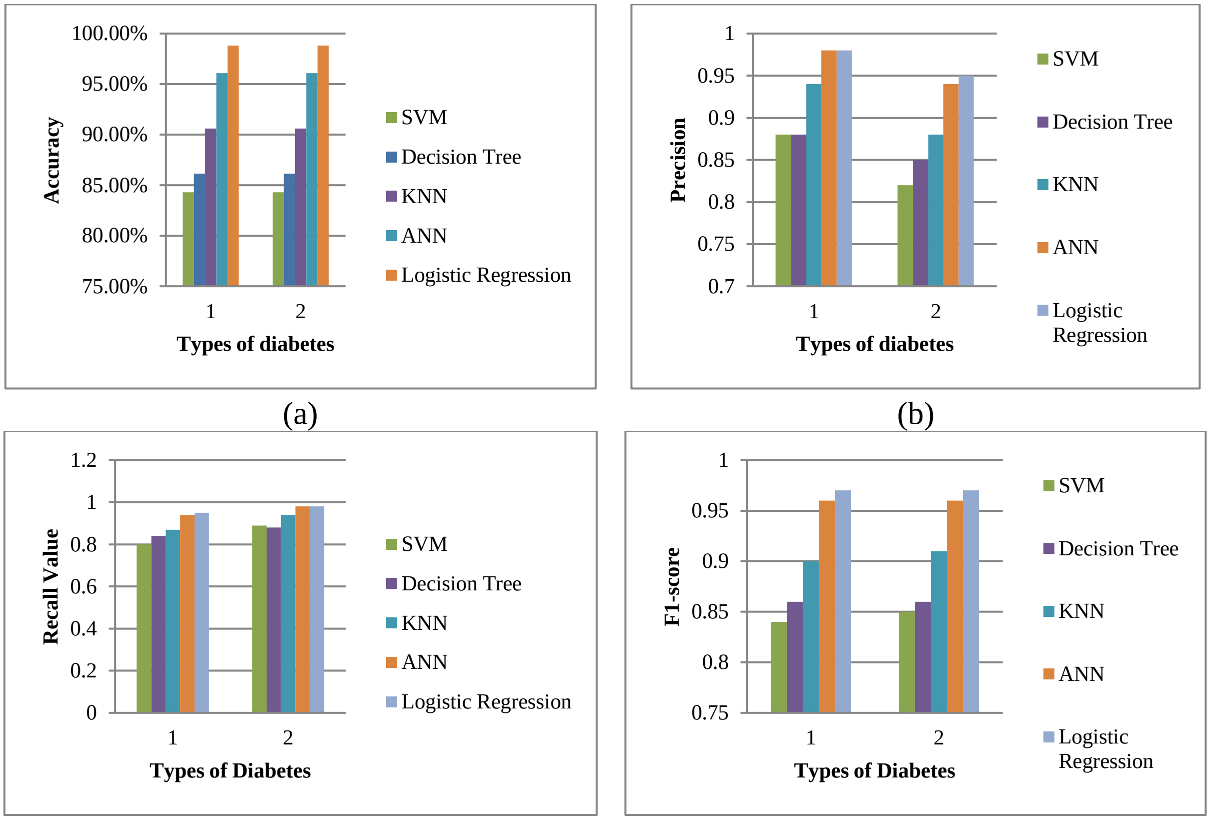
Algorithm	Class	n (truth)	n (classified)	Accuracy	Precision	Recall	F1 Score
SVM	1	1000	912	84.3%	0.88	0.80	0.84
	2	1000	1088	84.3%	0.82	0.89	0.85
Decision Tree	1	1000	961	86.15%	0.88	0.84	0.86
	2	1000	1039	86.15%	0.85	0.88	0.86
KNN	1	1000	930	90.6%	0.94	0.87	0.90
	2	1000	1070	90.6%	0.88	0.94	0.91
ANN	1	1000	960	96.1%	0.98	0.94	0.96
	2	1000	1040	96.1%	0.94	0.98	0.96
Logistic Regression	1	1000	972	98.8%	0.98	0.95	0.97
	2	1000	1028	98.8%	0.95	0.98	0.97

VIII. Comparison Of Different Machine Learning

An overview of the accuracy parameters for (SVM), Decision Trees, (KNN), and (ANN) algorithms is shown in Table 6. These algorithms are used to predict instances of type 1 along with type 2 diabetes. With a precision of 0.88, recall of 0.80, and F1 score of 0.84, (SVM) has an accuracy of 84.3% when it comes to its ability to predict instances of type 1 diabetes. With a precision of 0.82, recall of 0.89, along with F1 score of 0.85, the accuracy of predicting instances of type 2 diabetes is similarly 84.3%. This is in line with the previous statement. With an accuracy rate of 86.15% for both type 1 and type 2 diabetes patients, decision trees demonstrate a somewhat greater level of preciseness. The accuracy, recall, and F1 ratings for both kinds of instances are similarly impressive, which indicates that the performance is strong across the board. The accuracy of KNN is considerably higher, especially when it comes to forecasting instances of type 1 diabetes, with an accuracy of 90.6%. The algorithm is effective in properly identifying diabetic patients, as seen by the fact that the accuracy, recall, and F1 scores for both categories of cases are very high. The (ANN) emerges as the most accurate method, with an astonishing accuracy of 96.1% for treating patients of type 1 and type 2 diabetes. With accuracy, recall, and F1 scores that are regularly higher than 0.94, (ANN) demonstrate outstanding performance in the categorization of diabetes patients. The accuracy parameters accentuate the varied degrees of performance across the algorithms, with ANN demonstrating the greatest accuracy and KNN closely following suit. In general, the accuracy metrics highlight the differences in performance. These findings underline the significance of choosing the proper algorithm based on the particular needs and features of the dataset in order to obtain the highest possible level of prediction accuracy in the categorization of diabetes patients.

Table 6 Comparison of different ML

Parameters	**Type**	**SVM**	**Decision Tree**	**KNN**	**ANN**	**Logistic Regression**
Accuracy	1	84.3%	86.15%	90.6%	96.1%	98.8%
	2	84.3%	86.15%	90.6%	96.1%	98.8%
Precision	1	0.88	0.88	0.94	0.98	0.98
	2	0.82	0.85	0.88	0.94	0.95
Recall	1	0.80	0.84	0.87	0.94	0.95
	2	0.89	0.88	0.94	0.98	0.98
F1 Score	1	0.84	0.86	0.90	0.96	0.97
	2	0.85	0.86	0.91	0.96	0.97



(a) (b)

(c) (d)

Figure 3. Accuracy, Precision, Recall and F1 Score of Different algorithm to predict type 1 and type 2 diabetic

In figure 3 Accuracy, Precision, recall along with F1 Score is shown by using different algorithms that is SVM, KNN, ANN, Decision tree and get to know the best algorithm used for diabetic prediction. Focusing on situations with diabetes the bulk of the world's 422 million diabetics live in countries with low or medium economic levels. Deaths of 1.5 million a year are directly attributable to the illness. There has been a consistent increase in the prevalence and number of reported cases of diabetes throughout the last few decades.

IX. Conclusion

We are thinking of doing a survey of the most common methods currently utilized for categorization. Conclusion: Decision Tree yields an accuracy of 86%, whereas SVC and Perceptions achieve a minimum of 60% and 76%, respectively. The accuracy that is being provided by Bagging and KNN is 90%. Both Extra Trees and Random Forest have achieved 91% accuracy. All three classifiers—Gradient Boost, Ada Boost, and Gaussian NB—are equally accurate at 93%. With test accuracy rates of 96% and 94%, respectively, Logistic Regression and LDA have produced high-quality results. We find that ANN outperforms SVM, Decision tree, and KNN in this simulation in terms of accuracy. Support vector machines (SVMs) use a high-dimensional feature space for data categorization when linear axis partitioning is not possible. The information is changed when the border between the two groups is determined so that the boundary may be visually represented as a hyper plane. Extremely high blood sugar levels are the hallmark of the chronic disease known as diabetes. It has been linked to several complex diseases, including cardiovascular disease, kidney disease, along with stroke, among many others. Using a variety of ML approaches, including DL models, stacking methods, and others; our work seeks to detect illnesses in their early stages. Three stacking-based models in case of diabetes ailment classification are shown in this study by combining simulated data, additional data gathered from a local healthcare institution, and PIMA Indian diabetes dataset. We use predictions of many classification models utilizing the conventional and DNN stacking ensemble methodologies to increase classification accuracy along with resilience. During the evaluation phase, we used several classification techniques to evaluate our recommended model. After comparing all of the methods, we reflected on their performance results and chose the best one for classification approaches. In the past, researchers have utilized logistic regression to find the risk factors for diabetes based on the odds ratio and probability value. The authors use a wide variety of classifiers, such as NB, DT, AB, and RF, to forecast outcomes for individuals with diabetes. Twenty independent experiments were carried out, with each test using one of three distinct partitioning algorithms. Both the accuracy and the area under the curve are indicators of how well this classifier performs. Conventional study yielded an overall accuracy rating of 90.62% for ML systems. The K10 technique achieved an ACC of 94.25% and an AUC of 0.95 by combining an RF-based classifier with LR-based feature selection.

References

[1] A. Mahabub, “A robust voting approach for diabetes prediction using traditional machine learning techniques,” SN Applied Sciences, vol. 1, no. 12. Springer Science and Business Media LLC, Nov. 25, 2019. doi: 10.1007/s42452-019-1759-7.

[2] P. Thanh Noi and M. Kappas, “Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery,” Sensors, vol. 18, no. 2. MDPI AG, p. 18, Dec. 22, 2017. doi: 10.3390/s18010018.

[3] M. Alehegn, R. R. Joshi, and P. Mulay, “Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach,” Int. J. Sci. Technol. Res., vol. 8, no. 9, pp. 1346–1354, 2019

[4] M. A. Afrianto and M. Wasesa, “Booking Prediction Models for Peer-to-peer Accommodation Listings using Logistics Regression, Decision Tree, K-Nearest Neighbor, and Random Forest Classifiers,” Journal of Information Systems Engineering and Business Intelligence, vol. 6, no. 2. Universitas Airlangga, p. 123, Oct. 27, 2020. doi: 10.20473/jisebi.6.2.123-132.

[5] A. S. Agnal and E. Saraswathi, “Analyzing Diabetic Data Using Naive-Bayes Classifier,” Eur. J. Mol. Clin. Med., vol. 7, no. 4, pp. 2687–2698, 2020.

[6] N. I. Alghurair and M. A. Mezher, “Generic Frameworks for Svm , Ann , Lgbm , and Lr Algorithms,” Int. J. Comput. Sci. Mob. Comput., vol. 9, no. 6, pp. 132–140, 2020, [Online]. Available: https://www.academia.edu/download/63787039/V9I6202035.pdf

[7] E. S. Almutairi and M. F. Abbod, “Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia,” Modelling, vol. 4, no. 1. MDPI AG, pp. 37–55, Jan. 25, 2023. doi: 10.3390/modelling4010004.

[8] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” Decision Analytics Journal, vol. 3. Elsevier BV, p. 100071, Jun. 2022. doi: 10.1016/j.dajour.2022.100071.

[9] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” Journal of Applied Science and Technology Trends, vol. 2, no. 01. Interdisciplinary Publishing Academia, pp. 20–28, Mar. 24, 2021. doi: 10.38094/jastt20165.

[10] B. K. Kiranashree, V. Ambika, and A. D. Radhika, “Analysis on Machine Learning Techniques for Stress Detection among Employees,” Asian Journal of Computer Science and Technology, vol. 10, no. 1. The Research Publication, pp. 35–37, May 05, 2021. doi: 10.51983/ajcst-2021.10.1.2698.

[11] S. Kumari and A. Upadhaya, “Investigating Role of SVM, Decision Tree, KNN, ANN in Classification of Diabetic Patient Dataset,” Artificial Intelligence: Theory and Applications. Springer Nature Singapore, pp. 431–442, 2024. doi: 10.1007/978-981-99-8479-4_32.

[12] D. F. M. Mohideen, J. S. S. Raj, and R. S. P. Raj, “Regression Imputation and Optimized Gaussian Naïve Bayes Algorithm for an Enhanced Diabetes Mellitus Prediction Model,” Brazilian Archives of Biology and Technology, vol. 64. FapUNIFESP (SciELO), 2021. doi: 10.1590/1678-4324-2021210181.

[13] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” Procedia Computer Science, vol. 165. Elsevier BV, pp. 292–299, 2019. doi: 10.1016/j.procs.2020.01.047

[14] N. A. Noori and A. A. Yassin, “A Comparative Analysis for Diabetic Prediction Based on Machine Learning Techniques,” J. Basrah Res., vol. 47, no. 1, pp. 180–190, 2021, [Online]. Available: https://www.iasj.net/iasj/download/a371daadb33b96fd

[15] Ashwini Pathak and Sakshi Pathak, "Study on Decision Tree and KNN Algorithm for Intrusion Detection System," International Journal of Engineering Research and, vol. V9, no. 05. ESRSA Publications Pvt. Ltd., May 18, 2020. doi: 10.17577/ijertv9is050303.

[16] P. Rahman, A. Rifat, MD. IftehadAmjad Chy, M. Monirujjaman Khan, M. Masud, and S. Aljahdali, "Machine Learning and Artificial Neural Network for Predicting Heart Failure Risk," Computer Systems Science and Engineering, vol. 44, no. 1. Computers, Materials and Continua (Tech Science Press), pp. 757–775, 2023. doi: 10.32604/csse.2023.021469.

[17] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," Augmented Human Research, vol. 5, no. 1. Springer Science and Business Media LLC, Mar. 05, 2020. doi: 10.1007/s41133-020-00032-0.

[18] E. Zhu, Y. Ju, Z. Chen, F. Liu, and X. Fang, "DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features," Applied Soft Computing, vol. 95. Elsevier BV, p. 106505, Oct. 2020. doi: 10.1016/j.asoc.2020.106505.

[19] Z. Shaukat et al., "Revolutionizing Diabetes Diagnosis: Machine Learning Techniques Unleashed," Healthcare, vol. 11, no. 21. MDPI AG, p. 2864, Oct. 31, 2023. doi: 10.3390/healthcare11212864.

[20] A. A. Alhussan et al., "Classification of Diabetes Using Feature Selection and Hybrid Al-Biruni Earth Radius and Dipper Throated Optimization," Diagnostics, vol. 13, no. 12. MDPI AG, p. 2038, Jun. 12, 2023. doi: 10.3390/diagnostics13122038.

[21] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," International Journal of Cognitive Computing in Engineering, vol. 2. Elsevier BV, pp. 229–241, Jun. 2021. doi: 10.1016/j.ijcce.2021.12.001.

[22] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," BMC Bioinformatics, vol. 24, no. 1. Springer Science and Business Media LLC, Nov. 13, 2023. doi: 10.1186/s12859-023-05467-x.

[23] Md. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," Computer Methods and Programs in Biomedicine Update, vol. 4. Elsevier BV, p. 100118, 2023. doi: 10.1016/j.cmpbup.2023.100118.

[24] H. E. Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," Journal of ICT Standardization. River Publishers, May 21, 2022. doi: 10.13052/jicts2245-800x.10212.

[25] Kamal, A. Sharma, and D. Kumar, "Optimized Ensembled Model to Predict Diabetes Using Machine Learning," Optimized Predictive Models in Healthcare Using Machine Learning. Wiley, pp. 173–194, Feb. 07, 2024. doi: 10.1002/9781394175376.ch11.