# EFFICIENT DIABETES DISEASE PREDICTION: A MACHINE LEARNING AND PREPROCESSING FRAMEWORK

**Parag Jain**
**Research Scholar**
**Shobhit Institute of Engineering & Technology,**
**(Deemed to be University), Meerut**

**Nidhi Tyagi**
**Professor, CSE**
**Shobhit Institute of Engineering & Technology,**
**(Deemed to be University), Meerut**

**Birendra Kumar Sharma**
**Professor, MCA**
**Ajay Kumar Garg**
**Engineering College,**
**Ghaziabad**

## ABSTRACT

Modern computational intelligence is integrated into medical systems. Machine learning algorithms are used to find biomarkers for survival analysis based on the patient's specific health circumstances and to forecast the beginning and recurrence of the illness. Given the sharp rise in the number of diabetic patients across all age categories, early illness prediction is crucial. For medical professionals, figuring out the fundamental causes of diabetes at an early stage has become a bit challenging task. The volume of diabetic patient data is growing day by day, which has made it necessary to use effective machine learning algorithms. These algorithms identify patients' critical circumstances by learning from the trends in the underlying data. This research presents a cloud-based Internet of Things (IoT) platform for diabetes prediction. It uses sensors in wearable smart devices as a network of interconnectedIoTdevices for ongoing blood glucose data collecting and monitoring. The data is transferred to a cloud environment for storage, where an ensemble model is utilized to forecast diabetes in patients. Eight Ensemble models are used in the experiment, and two of the five possible machine learning techniques are paired. Using the "Pima Indians Diabetes" data set, the Decision Tree and Neural Network ensemble model yielded the greatest accuracy of 93.56%.

**Keywords:** Diabetes prediction, ensembled models, Machine learning, Neural Network, IoT

## INTRODUCTION

The number of people with diabetes and their mortality rate are on the rise globally, according to recent studies conducted by the World Health Organization (WHO). The WHO projected that by 2030, diabetes will rank as the sixth leading cause of death based on these trends [1]. One of the illnesses with the fastest global rise is diabetes. Diabetes is characterized as a group of metabolic diseases that raise blood glucose levels in people. The following are the two fundamental causes of elevated glucose levels: (1) the body's incapacity to generateenough insulin, and (2) the body's cells' improper response to insulin [2]. The hormone that is released by the pancreas and aids in controlling blood sugar is called insulin. The recommended range for blood sugar is 70–120 mg/dl, or 3.6–6.9 mmol/l [3]. Hypoglycemia is defined as glucose concentrations less than 50 mg/dl, which can cause increased thirst, perspiration, seizures, and diabetic coma. One job that is clinically relevant in the management of diabetes is hypoglycemic prediction. Preventive measures must be implemented and hypoglycemia should be anticipated well in advance due to its dangerous implications, which include coma and seizures. Higher glucose concentrations (>200 mg/dl) have been linked to hyperglycemia, which can cause nephropathy, neuropathy, and diabetic retinopathy as well as other long-term vascular problems. As a result, monitoring is required to effectively control glucose levels and improve overall quality of life. There are three forms of diabetes: Type-1, Type-2, and gestational diabetes. When the beta cells in the pancreas that produce insulin are destroyed by the immune system, type 1 diabetes results. Ten percent or so of people have Type 1 diabetes. Even if it is hard to avoid, the body can be treated by receiving insulin from an external source. On the other hand, Type-2 diabetes results from improper utilization of the insulin produced by the pancreas. Type-2 diabetes, which affects 90% of cases, is more frequent in those over 45 [4]. Patients with Type-2 diabetes have a two to four time’s higher risk of developing heart disease [5]. Gestational diabetes is the name for diabetes that affects women when they are pregnant. Blood glucometers measure diabetes at certain intervals. However, continuous glucose monitoring devices are used to assess diabetes continually. These devices offer a minimally intrusive way to record the patient's current glycemic level. Diabetes that is not detected in a timely manner has an impact on most bodily organs, including the kidneys, eyes, heart, nerves, and so forth. As a result, it's critical to make an early and precise diabetes prognosis. The diagnosis and interpretation of diabetes, together with the proper analysis of data, become critical issues when tackling them as a machine learning (ML) classification problem. For this reason, the effective prediction of diabetes is much appreciated when computational intelligence is applied.

Emerging technologies in today's world include artificial intelligence (AI), machine learning (ML), deep learning (DL), big data, and the Internet of Things[6, 7]. Patients mayquickly determine their condition in the early stages with the use of machine learning (ML), which will also assist researchers in their future work. It is applicable to situations involving both regression and classification. Since diabetes prediction is a classification problem, people can be categorized into several groups based on whether they have diabetes or not [8]. A variety of machine learning approaches may be applied to analyze and summarize the data into insightful knowledge from different angles. Preprocessing the dataset, selecting and extracting features, training and testing, and further assessment are some of the procedures involved in machine learning. Many types of data are gathered, including text, sensor, and clinical data [9]. These data are produced by distinct wearable devices, which produce data that is primarily unprocessed. Preprocessing is required to transform this data into a comprehensible format. For effective diabetes prediction, a variety of machine learning techniques, such as Random Forest (RF), Decision Tree (DT), Neural Network (NN), and Vector Machine (SVM), can be used [10]. After being trained, these models are verified to see if they are functioning properly using the test dataset. Researchers [11] have been putting a lot of effort into predicting diabetes in the past, but the results are insufficient. Therefore, it is necessary to provide additional methods for precise and effective prediction.

One of the most common chronic diseases that affects people worldwide in all age categories is diabetes mellitus. It exhibits either no symptoms at all or very minor signs for a long time before being discovered. It has been shown that undiagnosed diabetes can damage other essential organs in the body. For several ages, researchers have studied DM.The term "diabetes" was initially used by the 5th-century physician Aristaeus, who described the illness as a "melting down of limbs and flesh into urine."The sweet, honey-like flavor of polyureic patients' urine that attracted ants and other insects was identified by Indian physicians in the fifth century BC; nevertheless, the word Mellitus Latin for "honey" was added in the seventeenth century.Diabetes mellitus (DM) is a chronic, non-communicable condition characterized by partial or total insulin insufficiency.Although the causes of diabetes might vary greatly, they are always related to either the body's cells not responding appropriately to insulin or the pancreas' inability to secrete enough insulin at some time over the course of the illness.Thus, in order to save lives, an early forecast is necessary. Researchers have attempted to predict diabetes early using a variety of machine learning methods, such as SVM, RF, NN, KNN, and DT, but the results have not been adequate. As a result, sophisticated methods are needed to more accurately forecast diabetes. Development of diabetes mellitus is strongly linked to the main problems in the metabolism of fat, protein, glucose, and carbohydrates.Additionally, it is linked to a lower life expectancy, significant morbidity from diabetes-related microvascular issues, macrovascular issues (such as peripheral vascular disease, stroke, and heart disease), infections, and other consequences (such as COVID-19, nonalcoholic fatty liver disease, dental disease, dyslipidemia, and psychosocial issues), as well as a lower quality of life.Diabetes is one of the major social and public health issues of our day due to its prevalence, special features, and the frequency of other illnesses that often co-occur with it.

In most of the prior research focus is to improve the model for better accuracy, there is no significant research to pre-process the dataset to get the better results. In this paper, we are proposing to pre-process the dataset before train the model and predict the results. Pre-process is mainly focused on to remove the duplicate records, fill the missing values and select the correct features, which are contributing more to predict the result for better accuracy and remove the unnecessary features.

## . LITERATURE REVIEW

These days, a number of novel models such as machine learning system classifiers are employed in the early diagnosis of diabetes. To get the biomedical statistics, several data mining techniques are employed for the disease's prognosis. The term "data mining technique" refers to the process of gathering and analyzing data from several databases in order to extract significant portions that are needed for efficient analysis. Predictive analytics is being applied in the field of diabetes to aid in diagnosis, prognosis, self-management, and prevention. In the current trend, research analysis aids in the prediction of diabetes, which is important in cases of high mortality and high morbidity as well as the compilation and prevention of diseases. Worldwide, data scientists and medical professionals view diabetes prediction as a challenge. ML algorithms were proposed by Quan Zou et al. [12] for the diagnosis of Type 2 Diabetes. A sample of 68,994 healthy patients was obtained over five iterations of data collection. The outcomes demonstrated that, with an accuracy of 80%, the suggested strategy was effective. ML methods, such as AdaBoost with Decision Stump, DT, SVM, and Naive Bayes (NB), are used by Veena et al. [13] to diagnose and prognosticate diabetes mellitus. The experiment was carried out using a sample that was obtained from the PIMA dataset. The approach to reliably diagnose diabetes mellitus was developed by Panwar et al. [14] in light of new pre-processing methods and the K-closest neighbor classifier. The author overlooked the diabetic pedigree function and diastolic blood pressure, using the remaining six variables for categorization. The evaluation parameters were optimized by adjusting a few factors.The author optimized the following parameters: distance function, leaf size, and number of neighbors for K-nearest neighbor. Sowjanya et al. [15] used the Android/portable application as a means of responding to the lack of awareness on diabetes. The obtained data is arranged using four machine learning algorithms: J48, naive Bayes, support vector machines, and multilayer perceptron's. A healthcare system utilizing decision trees and k-nearest neighbor to forecast diabetes was proposed by Hashi et al. [16]. Using the PIMA diabetes dataset, the model was trained to achieve 90.43% accuracy. The Pima Indian Diabetic Data Set serves as an instructive index for much of the relevant literature. Early diabetes prediction is crucial since it lessens the deadly consequences of the disease. The Writing Survey of Diabetes Assumptions shows that a single diabetes identification technique is not a very sophisticated way to find diabetes at an early stage. Anand [17] presented a model that takes everyday living activities into account to forecast a patient's likelihood of developing diabetes. Using the PIMA dataset, the classification and regression tree technique yields a 75% accuracy rate. Jakhmola [18] presented a model to predict a person's likelihood of having diabetes using multiple regression analysis and controlled binning. With the PIMA dataset, the accuracy of the model is 77.85%.Jarullah [19] used pre-processing methods and the J48 classifier to create a decision tree model [20]. With the PIMA diabetes dataset, the accuracy of the model is 78.17%. KSVM, a hybrid model utilizing the feature selection approach, was introduced by Hamza et al. [21]. Prag et al. [22] has done the comparative study for existing models and accuracy they have found in that modal and conclude that preprocessing of dataset is also required and play very critical role in model implementation.

For the purpose of detecting different diseases, several automated systems are created in the fields of data mining, machine learning, fuzzy logic, and neural networks [23]. The Oracle Data Miner (ODM) tool, which uses the Support Vector Machine method, was utilized to forecast the course of treatment for both young and senior diabetes patients. According to Abdullah Aljumah A, et al. [24], elderly diabetes patients should begin therapy as soon as feasible, whereas young diabetic patients should wait to avoid adverse effects. When Joseph L. et al. [25] used Classification and Regression Tree (CART) to predict the onset of diabetes, they found that younger persons were more susceptible to the condition than older ones. After pre-processing, Santhanam et al. [26] employed the K-means method to eliminate noise from the data. The study demonstrated that the K-means algorithm [27] outperformed the Support Vector Machine in the diagnosis of diabetes in Pima pregnant women. In order to maximize the AUC parameter, the author conducted extensive tests using several pre-processing technique groups and machine learning algorithms. As a baseline model for the evaluation of the projected ensemble classifier, the method that produced the best results is suggested. Soft weighted voting is used to combine the Adaboost and gradient boost machine learning algorithms. Because the AUC is balanced towards the class distribution, it is chosen as the model's weight for voting. In order to diagnose diabetes using a multi-model evolutionary algorithm, Catalin Stoean et al. [28] suggested the Elitist Generation Genetic Chromodynamics algorithm (EGGC). The current algorithm incorporates If-Then rules by utilizing the idea of evolution. The rules that are acquired have great significance since they not only offer the outcomes but also the rationale that guides the decision-making process. During the last ten years, there has been a lot of research being done on the design of prediction models for diabetes diagnosis. Artificial neural networks (ANNs) and classification algorithms provide the foundation of the majority of models found in literature. Below are a few of the research articles that were examined for this study:Diabetes has been predicted using Decision tree (J48) and Naïve Bayes algorithms by Aiswarya Iyer et al. [29]. The Pima Indian Diabetes dataset was utilized, and the WEKA tool was employed to apply it. They discovered that the Naive Bayes algorithm predicted diabetes with 79.56% more accuracy than another.SVM with Radial Basis Function Kernal was employed by V. Anuja Kumari and R. Chitra [30] for the categorization of diabetes. For implementation, PYTHON, R2010a was utilized. They discovered a 78% accuracy rate. Data mining classification algorithms including CART, J48, and NBTree were used by Satheeskumar and Gayathri [31] to analyze adult-onset diabetes. To put these algorithms into practice, they employed the WEKA tool. When compared to other algorithms, they discovered that the J48 method had an accuracy rate of 80%. Using 18 risk indicators. Data mining was employed by V. Kumar and L. Velide [32] to predict and treat diabetes. They employed JRip, DT, NN, J48 (4.5), and Naïve Bayes as their approaches. They implemented it using the WEKA tool. They achieved a J48 algorithm accuracy rating of 68.5%. Data mining techniques were offered by Srideivanai Nagarajan and R.M. Chandrasekaran [33] as a way to enhance the diagnosis of gestational diabetes. Additionally, they examine the effectiveness of ID3, Naïve Bayes, C4.5, and Random Tree algorithms, which are all related to supervise learning.They made advantage of the pregnant women's data collection.According to their findings, the Random tree performed the best, having the lowest mistake rate and more accuracy.

## PRE-PROCESSING TECHNIQUES

A large number of rows and columns is referred to as data, but it can also take the shape of pictures, movies, tables, music, etc. The data must be converted into a format that can be utilized to train the machine in order for computers to use it for decision-making purposes. This means that the data cannot be used directly in the forms of free text, photos, audio, or video. Data pre-processing is the process of converting data into a format that makes it simple for a machine to parse. Following data pre-processing, the machine learning algorithm can readily analyze the input. Many stages are taken to pre-process the dataset; however, not every step needs to be used for every issue. The number of steps that apply in a given scenario greatly depends on the data that we are dealing with. Of all the pre-processing techniques, cleaning of data, normalization and feature selection are the most often applied techniques.

Process a huge size of data is quite complex and if data is not in good condition, it impacts the accuracy of prediction algorithm as well. Raw data should be cleaned before using the data for algorithm. Cleaning of data includes, removing duplicate values, missing filled values.

Feature selection is most important part of pre-processing of data. Dataset may contain thousands of features, but it is not necessary that every feature is contributing in predicting the result or it is contributing very less, which is negligible for large size of dataset.

## PROPOSED FRAMEWORK

The literature has presented a number of methods for predicting the early identification of diabetes. The majority of these methods rely on antiquated machine learning frameworks, which are ineffective in relations of accuracy. In this paper, we have presented an ensemble framework to preprocess the dataset for diabetes prediction using the data set of a featured diabetic patient. For better accuracy, data clean-up is required, which includes the removing the duplicate values and filling the missing values. To eliminate unnecessary features, feature selection is required. There are three methods for selecting features: the filter technique, the wrapper method, and the embedding method.The four fundamental phases of a feature selection process are subset creation, subset estimation, stopping condition, and result confirmation. For feature selection, we have used correlation of features in output. Higher is the correlation for each feature with the output, needs to be considered. Figure 1 represents the process to clean the dataset and select the most corelatated features, which contributes more in the final ouutcome.

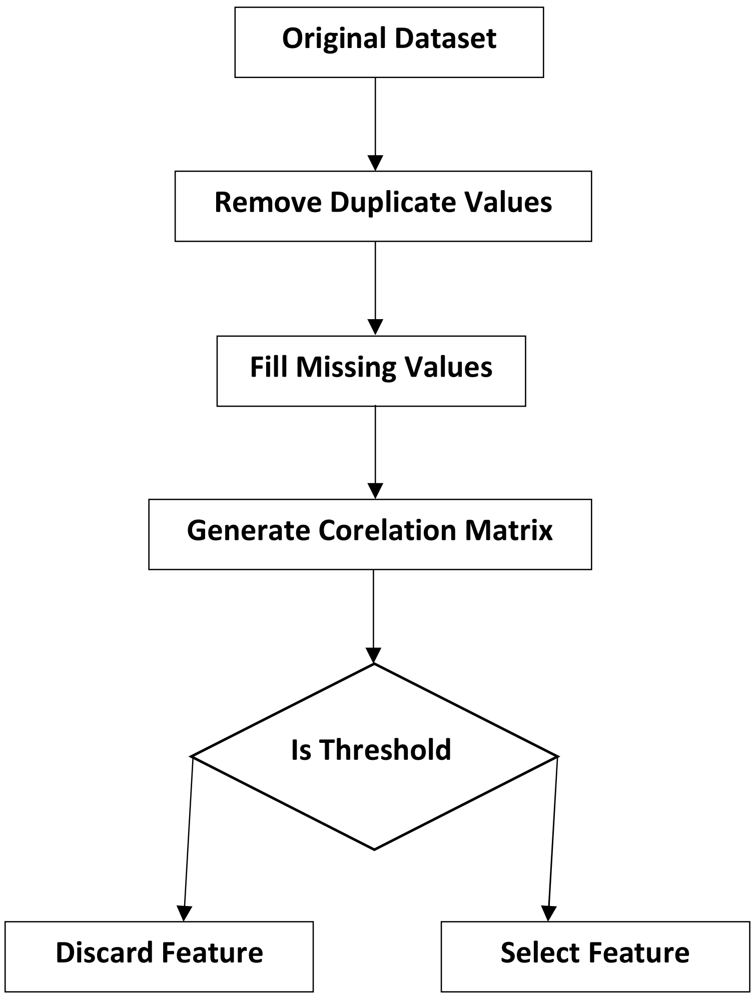

**Figure 1:**Process Flow Diagram for Preprocessing of dataset

### 1. Dataset

In the proposed framework, we have used Pima Indian Diabetes dataset, which includes 768 records and it has eight features and one column for outcome. This dataset is well known used for machine learning and statistical analysis tasks related to diabetes prediction. It originates from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of medical data collected from a population of women of Pima Indiana Heritage. The primary objective of this dataset is to predict whether a patient has diabetes based on certain diagnostic measurement.

#### 1.1 Features

**Pregnancies –** Number of times pregnant

**Glucose –** Plasma glucose concentration 2 hours in an oral glucose tolerance test

**Blood Pressure –** Diastolic blood pressure (mm Hg)

**Skin Thickness –** Triceps skin fold thickness (mm)

**Insulin –** 2-Hour's serum Insulin (mu U/ml)

**BMI –** Body Mass Index (Weight in kg/ (height in m) ^2)

**Diabetes Pedigree Function –** A function which score likelihood of diabetes based on family history

**Age –** Age (Years)

**Outcome –** Class Variable (0 or 1), 0 means no Diabetes, 1 mean Diabetes.

### 2. Cleaning Dataset

Cleaning of dataset includes removing duplicates values and filling the missing values. Missing values in a dataset can significantly affects the performance of machine learning models, leading to incorrect predictions and reduced model accuracy. In the Pima Indian Diabetes Dataset, missing values are represented by zeros in certain columns where such values are physically impossible. These columns include 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', and 'BMI'. In our proposed pre-processing technique, we have filled the missing values with the mean value for better accuracy.

Duplicate values in a dataset can introduce bias and skew the results of data analysis and machine learning models. It is essential to detect and remove duplicate records to ensure data quality and reliability.

### 3. Feature Selection

Feature selection is an essential process in machine learning process that involves selecting a subset of relevant features for model training. Using highly correlated features can lead to multicollinearity, which can affect the performance of the machine leaning model. Correlation measures the strength and direction of the linear relationship between two variables. The correlation coefficient ranges from -1 to 1.

- A correlation of 1 indicates the prefect positive linear relationship.

- A correlation of -1 indicates the prefect negative linear relationship.
- A correlation of 0 indicates no linear relationship.

Figure 2 shows the Correlation Matrix for Pima Indian Diabetes dataset.

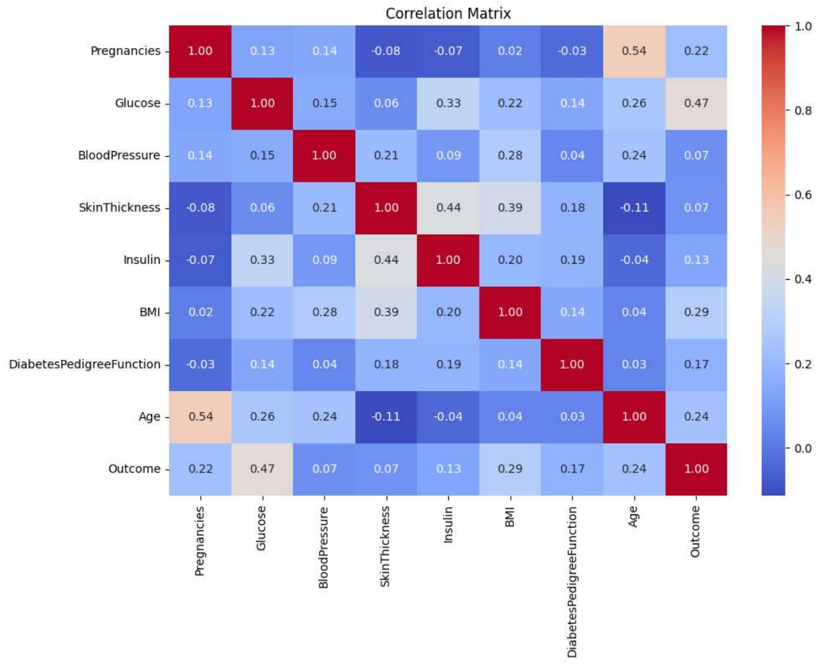

**Figure 2:**Correlation Matrix

From the Correlation Matrix mentioned in Figure 1, it is clearly visible that few of the features are very less impacting or negligible impacting the outcome. For our study we have dropped the features, which are have correlation value 0.2 or less and considered the following feature in prediction.

- **Pregnancies**
- **Glucose**
- **BMI**
- **Age**

### 4. Methodology

#### a) Training Phase

In this phase, we have taken the output of prediction layer i.e. pre-processed data and developed an ensemble machine learning model using Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB) and Support Vector Machine (SVM). All these models are trained individually in the initial phase of model building. After the initial training, the final prediction output is fetched. Pseudo code for the training phase is given as algorithm 1.

**Algorithm 1. CI-DPF**

```
Input: PIMA Indian Diabetes dataset, G= {LR, KNN, NB, SVM}
Output: Ensemble Prediction Model(M)
Step 1: Data cleansing and filtering: removing duplicate entries and missing values.
Step 2. Feature Selection // Correlation Matrix
Step 3: Individual Model Training (X) // X ∈G
Step 6: Ensemble Modelling(M)
```

#### b) Prediction Phase

In order to assess the efficacy and robustness of the intended framework, we have predicted test samples from the PIMA Indian Diabetes Dataset in this phase. The Ensemble Prediction Model (M), which is the result of the training step, serves as the input for this stage. The algorithm 2 provides the pseudo code for the prediction step.

**Algorithm 2. CI-DPF(Prediction)**

```
Input: Test Sample(S), Ensemble Model(M)
Output: Classified Sample (O)
Step 1. O = M(S). // Sample Prediction.
Return O // Predicted Output.
```

## RESULTS AND DISCUSSION

Healthcare systems provide patients with specialized services in a wide range of areas to help them integrate into their daily routines. One of the most important and serious issues facing the medical field is diabetes mellitus.One of the most important decision-making techniques in the real-world situations of today is classification. The main objective is to classify the data as either non-diabetic or diabetic and improve the accuracy of the classification. The main goal of machine learning in diabetes diagnosis is to identify patterns in the provided diabetes dataset.The medical field has long relied on machine learning as a reliable, supporting, and ever-evolving technology.This work focuses on using machine learning classifiers to identify diabetes and identify patient categories based on specific medical information. Finding the limitations of recommended works in the area of machine learning classifiers for diabetic patients' treatment regimens is helpful. The diagnosis of diabetes is an expanding field of study. A range of medical tests are needed to diagnose a particular ailment in science and medicine. The diagnosis is dependent on the experience of the physician; a less skilled practitioner may identify a problem incorrectly. Purpose of this study was to pre-process the dataset and select the most corelated features for better accuracy and efficiency.

### 5. Performance Parameters

The suggested paradigm is validated using the performance metrics of Accuracy, Sensitivity, and Specificity. Here, TN (number of correctly predicted negative instances), FP (number of incorrectly predicted positive instances), FN (number of incorrectly predicted negative instances), and TP (number of correctly predicted positive instances) are recorded.

i. $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

ii. $$Sensitivity = \frac{TP}{TP+FN}$$

iii. $$Specificity = \frac{TN}{TN+FP}$$

Table1 demonstrates that for the diabetes dataset, Logistic Regression yields the best accuracy of 78.35 and K-Nearest Neighbor the lowest accuracy of 74.46 without pre-processing the dataset. The investigation leads to the conclusion that the Logistic Regression model outperforms the other models in terms of accuracy. We need to pre-process the dataset in order to increase the accuracy of the weaker models. The Table 1 outcome statistics are displayed in Figure 3.

**Table 1:**Existing Model Accuracy without Preprocessing the dataset

| S. No. | Model Name | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| 1 | Support Vector Machine | 76.62 | 58.02 | 86.67 |
| 2 | K-Nearest Neighbor | 74.46 | 54.32 | 85.33 |
| 3 | Naive Bayes | 76.19 | 64.20 | 82.67 |
| 4 | Logistic Regression | 78.35 | 60.49 | 88.00 |

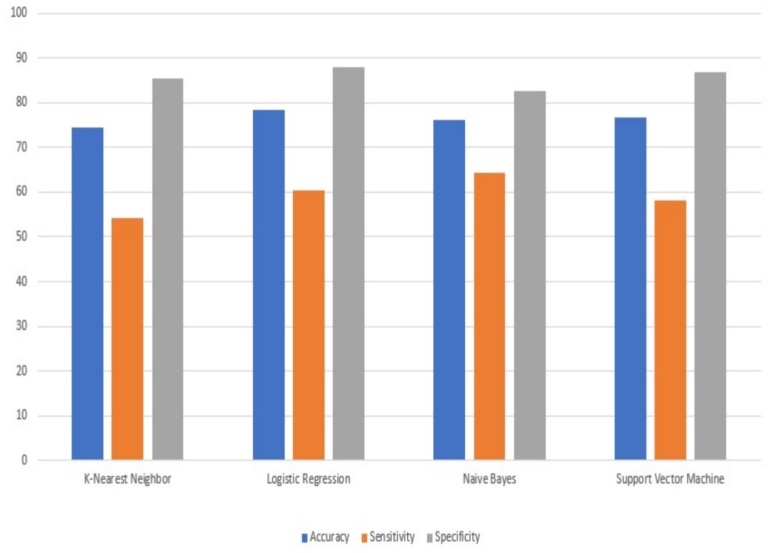

**Figure 3**: Performance without Preprocessing

As mentioned in the proposed framework for preprocessing, we have cleaned the dataset by removing the duplicate results and filling the missing values. Feature selection is one of the key factors in preprocessing. Table-2 is the outcome of algorithms with preprocessed dataset.

**Table 2:**Pre-Processed Dataset model results

| S. No. | Model Name | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| 1 | Support Vector Machine | 77.49 | 56.79 | 88.67 |
| 2 | K-Nearest Neighbor | 77.06 | 62.96 | 84.67 |
| 3 | Naive Bayes | 74.46 | 56.79 | 84.00 |
| 4 | Logistic Regression | 77.06 | 58.02 | 87.33 |

As shown in Table 2, after preprocessing the dataset, existing algorithm accuracy has been increased and Support Vector Machine outperforms as compare to other models with the 77.49% accuracy. K-Nearest Neighbor is showing lot of improvement, as with raw dataset accuracy was 74.46% and after preprocessing the dataset, accuracy increased to 77.06%. The Table 2 outcome statistics are displayed in Figure 4.

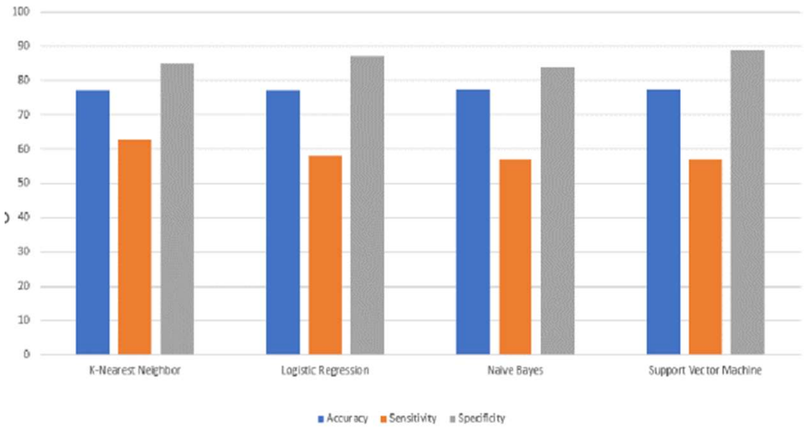

**Figure 4:** Performance parameters based on Table 2

## Conclusion and Future Work

To improve the accuracy of prediction for Diabetes preprocessing of dataset has been proposed. We have used the PIMA Indian Diabetes Dataset and seen that after preprocessing of dataset, few models have performed better as compare to raw dataset. Support Vector Machine outperforms as compare to other models with the 77.49% accuracy. K-Nearest Neighbor is showing lot of improvement, as with raw dataset accuracy was 74.46% and after preprocessing the dataset, accuracy increased to 77.06%. Form the proposed methodology, it is concluded that cleaning of data is required before processing the dataset. Also feature selection is the critical part for implementing any model to improve the accuracy.

In future, need to do more work in feature selection and apart of dropping the features from dataset, it is also required to identify and add new features, which may contribute to improve the accuracy and efficiency of the model. To track the patients' blood sugar levels, a cloud-based, Internet of Things-based framework for diabetes prediction can also be implemented.

## REFERENCES

[1]. Parampreet Kaur, Neha Sharma, Ashima Singh, and Bob Gill. 2019. CI-DPF: A cloud IoT based framework for diabetes prediction. InProceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON'18), 654–660. DOI:https://doi.org/10.1109/IEMCON.2018.8614775

[2]. Chitra Jegan, V. Anuja Kumari, and R. Chitra. 2018. Classification of diabetes disease using support vectormachine. Int. J. Eng. Res. Appl.3, 2 (2018), 1797–1801. Retrieved from https://www.researchgate.net/publication/320395340

[3]. Diseases Conditions. Retrievedfrom https://www.mayoclinic.org/diseases conditions/diabetes/diagnosistreatment/drc-20371451

[4]. Chatrati, S. P., Hossain, G., Goyal, A., Bhan, A., Bhattacharya, S., Gaurav, D., & Tiwari, S. M. (2022). Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University-Computer and Information Sciences*, *34*(3), 862-870.

[5]. Shivsharanr, N., & Ganorkar, S. (2021). Predicting Severity of Diabetic Retinopathy using Deep Learning Models. International Research Journal on Advanced Science Hub, 3(Special Issue ICEST 1S), 67-72.

[6]. Sosale, B., Aravind, S. R., Murthy, H., Narayana, S., Sharma, U., Gowda, S. G., & Naveenam, M. (2020). Simple, mobile-based artificial intelligence algorithm in the detection of diabetic retinopathy (SMART) study. *BMJ Open Diabetes Research & Care*, *8*(1).

[7]. Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, and Frank Schwartz. 2014. A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes. AAAI Workshop Technical Report WS-14-08 (2014), 35–39.

[8]. Ambika Choudhury and Deepak Gupta. 2019.Recent Developments in Machine Learning and Data Analytics. Springer Singapore.DOI:https://doi.org/10.1007/978-981-13-1280-9

[9]. Gagangeet Singh Aujla, Anish Jindal, Rajat Chaudhary, Neeraj Kumar, Sahil Vashist, Neeraj Sharma, and Mohammad S. Obaidat. 2019. DLRS: Deep learning-based recommender system for smart healthcare ecosystem. InProceedings of the IEEE International Conference on Communications. DOI:https://doi.org/10.1109/ICC.2019.8761416

[10]. Arwinder Dhillon, Ashima Singh 2019. Mach. Learn.Healthcare.8, (July 2019), 92–109.

[11]. Ahmed, A., Qayoum, A., & Mir, F. Q. (2019). Investigation of the thermal behavior of the natural insulation materials for low temperature regions. *Journal of Building Engineering*, *26*, 100849.

[12]. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. 2018. Predicting diabetes mellitus with machine learning techniques.Front. Genet.9, (2018) 1–10.DOI:https://doi.org/10.3389/fgene.2018.00515

[13]. V. Veena Vijayan and C. Anjali. 2016. Prediction and diagnosis of diabetes mellitus—A machine learning approach, InProceedings of the 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS'15), 122–127. DOI:https: //doi.org/10.1109/RAICS.2015.7488400

[14]. M. Panwar, A. Acharyya, R.A. Shafik, and D. Biswas, "Knearest neighbor based methodology for accurate diagnosis of diabetes mellitus". In Embedded Computing and System Design (ISED), pp. 132-136. IEEE,2016.

[15]. K. Sowjanya, A. Singhal, and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," In Advance Computing Conference (IACC), pp. 397-402. IEEE, 2015.

[16]. E. K. Hashi, M. S. U. Zaman, and M. R. Hasan, "An expert clinical decision support system to predict disease using classification techniques," 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 2017.

[17]. A. Anand and D. Shakti, "Prediction of diabetes based on personal lifestyle indicators," 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2015

[18]. S. Jakhmola and T. Pradhan, "A Computational Approach of Data Smoothening and Prediction of Diabetes Dataset," Proceedings of the Third International Symposium on Women in Computing and Informatics - WCI 15, 2015.

[19]. A. A. A. Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," 2011 International conference on innovations in information technology. New York, IEEE, 2011.

[20]. García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., & García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine*, *202*, 105968.

[21]. A. Hamza and H. Moetque, "Diabetes Disease Diagnosis Method based on Feature Extraction using KSVM," International Journal of Advanced Computer Science and Applications, vol. 8, no. 1, 2017.

[22]. Prag Jain and Nidhi Tyagi, "Diabetes Mellitus Prediction Algorithm – Comparative Study," Anvesak, vol. 53, no. 2, 2023.

[23]. Priyanka khare, Dr. Kavita Burse, "Feature Selection using Genetic Algorithm and Classification using Weka for Ovarian Cancer", International Journal of Computer Science and Information Technology, Vol. 7, No. 1, pp. 194-196, 2016.

[24]. Abdullah Aljumah A, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, “Application of data mining: Diabetes healthcare in young and old patients”, Journal of King Saud University – Computer and Information Sciences, Vol 25, pp. 127 – 136, 2013.

[25]. Joseph L. Breaulta B, Colin R. Goodall.C.D, Peter J. Fose B, “Data mining a diabetic data warehouse”, Artificial Intelligence in Medicine, Vol -26, pp- 37–54, 2002.

[26]. Santhanam T, Padmavathi. M. S, ”Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis”, Procedia Computer Science, Vol – 47, pp-76 – 83, 2015

[27]. D. Westari, A. Halim and M. Eng, Performa Comparison of the K-Means Method for Classification in Diabetes Patients Using Two Normalization Methods 04(01) (2021), 18-23. doi:10.47191/ijmra/v4-i1-03

[28]. Catalin Stoean, Ruxandra Stoean, Mike Preuss and D. Dumitrescu, “Diabetes Diagnosis through the Means of a Multimodal Evolutionary Algorithm”, Elsevier, pp.76-82, 2015

[29]. Aiswarya Iyer,S.Jeyalatha,RonakSumbaly,“Diagnosisofdiabetesusingclassificationminingtechniques”,(IJDKP), Vol.5, No.1, January 2015, pp. 1-14.

[30]. V. A. Kumari and R. Chitra, “Classification of Diabetes Disease using Support Vector Machine,” IJERA, Apr. 2013.

[31]. S. Kumar B and G. P, “Analysis of Adult-Onset Diabetes using Data mining Classification Algorithms,” IJMCS, vol. 2, no. 3, Jun. 2014.

[32]. V. Kumar and L. Velide, “A Data mining Approach for Prediction and Treatment Ofdiabetes Disease,” IJSIT, 2014.

[33]. S. Nagarajan and R.M.Chandrasekaran, ”Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes” in proceedings of International Journal of Current Research and academic Review, vol. 2,No. 10,pp. 91-98.