International Journal of Computing and Digital Systems

Enhancing Diabetes Prediction Using Ensemble Machine Learning Model

Dr. Aniket K. Shahade[1], Dr. Priyanka V. Deshmukh[1]

[1]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India
E-mail address: aniket.shahade@sitpune.edu.in, priyanka.deshmukh@ sitpune.edu.in

Abstract: Diabetes is a disease which is beyond cure and which has adverse effects on the health and hence has to be detected at an earlier time to avoid more damage to the body. This study aims at establishing the use of machine learning in the circumstances of diabetes prediction based on factors such as glucose levels, blood pressure, skin fold thickness, and insulin. The purpose of this study is to identify the potential of using machine learning techniques, such as Support Vector Machine (SVM), Logistic Regression and proposed Ensemble Model for the prediction of diabetes.

To this aim, in the current study, a dataset including the fundamental medical features of a general population of patients was employed. Regarding this, the data pre-processing was done with the view of handling missing data, data normalization and feature extraction in a view of enhancing the performance of the proposed model. All the models have been developed, and the data was split to perform k-fold cross validation to make the predictions more accurate.

From the evaluation metrics, it is evident that the proposed Ensemble Model is the most appropriate since it has a higher accuracy rate compared to the Support Vector Machine, Logistic Regression model. To compare the performance of each model the metrics used includes accuracy, precision, recall, F1-Score.

Therefore, the above analysis shows that the proposed Ensemble model is effective in the prediction of diabetes, and this is why there is the need to consider data mining in order to improve the health care delivery systems.

Keywords: Diabetes Prediction, Machine Learning, Ensemble Learning, Logistic Regression, Support Vector Machine, Random Forest, Medical Data Analysis, Predictive Modeling, Health Informatics, Diabetes Risk Factors

1. Introduction

1.1 Background on Diabetes

Diabetes: Prevalence and Impact

This aims at evaluating the effects of some chronic diseases and diabetes mellitus could be regarded as one of them as it affects millions of people within the entire world. Diabetes is also known diabetes is a disease in which there is compound high level of blood sugar and is as a result of insufficient insulin or the inability of internal body to use insulin produced. In the year 2019, the International Diabetes Federation estimated that the population affected by diabetes was 463 million and predicted that more people might be diagnosed with the disease in the future. It is a disease that is characterized with increased morbidity and mortality, has complications in the cardiorenal, nervous as well as ocular systems and more nurses globally through the effects it has on health care systems.

It also impacts the economy as a lot of money is spent on medical care, drugs and other related expenses in the management of diabetes and its complications. Therefore, apart from the financial aspect of the disease, diabetes has numerous effects on the health of patients as the treatment for the disease is life-long and the patients have to be cautious with their health all the time.

Importance of Early Diagnosis and Prediction

The early identification of the disease and the possibility of identifying the onset of diabetes are crucial in minimizing the effects of diabetes and in improving the life expectancy of the patients. This is because identifying people who may be at risk of developing the condition at some point in the future means that changes in diet and other lifestyles can be made in order to reduce the risk of getting diabetes. It also assists in controlling the disease, thus averting the occurrence of severe symptoms or complications of the disease, thus increasing the patient's life expectancy.

This is because there are enhanced techniques in the health sector and the growth in the big data systems for early diagnosis and prognosis. This paper therefore argues that Machine Learning and other forms of Artificial Intelligence are useful tools for predicting diabetes risks from large complex medical data sets. By using these technologies, the health care providers will be

E-mail: aniket.shahade@sitpune.edu.in

in a better position to diagnose the patient and treat them which will lead to the improvement of the services that the patient will be receiving from the health care providers.

1.2 Importance of Machine Learning in Healthcare

1.2.1 Applications of AI and ML in Medical

Diagnosis: An Overview

With the advancement of AI & ML, ventures have become very different and have contributed significantly to the healthcare fields in various countries. By so doing , the healthcare systems are in a position to handle sometimes very large volumes of data in the medical niche so that doctors can easily diagnose diseases correctly and in a faster manner . It could also be appropriate to use AI and ML algorithms because they can search for reflections of certain patterns and show correlations in very large data sets that any human practitioner might not even consider noticeable. These are as follows; in department of radiology an algorithm intervention for example used in differentiation of X-Ray and MRI images for lesions and in any area of predictive analytics for example use in epidemic and patient status prognosis.

In clinical practices, AI and ML are used in establishing algorithms helpful in diagnosing diseases like cancers, cardiovascular diseases, neurological disorders, among others. AI is a broad concept and NLP is one of the subsets and in it, diagnostic and descriptive texts which may include clinical notes or the electronic health records (EHRs) are assessed to gain important insights that can be used in decision-making processes about the treatment to be provided. Additionally, as personalised medicine states applying targeted treatments and diagnostic approaches depending on patient's genetic profile, the application of machine learning in analysis of genomes is rather beneficial.

Specific Relevance to Diabetes Prediction

In the context of pattern classification for outcome prediction in relation to affiliation to diabetes, therefore, machine learning has the following benefits. Diabetes is one of the chronic diseases that has numerous important global ontogenetic determinants, including genetic and behavioural factors and other various co-morbid conditions. Standard diagnostic techniques are also relatively effective but at the same time they cannot perform more sophisticated procedures such as a multivariate analysis and therefore may fail to discover the interactions between them. Still as for keeping these intricate relations, they do this quite well as soon as various ML algorithms are being applied.

In the examined medical data, such as information about the patients with glucose, pressure, and insulin levels, some algorithms let predicting diabetes with quite high likelihood. They are logistic regression analysis methods and support vector machines (SVM) and logistic regression analysis methods, random forests, support vector machines (SVM) that analyse these variables to arrive at the predictive information. These models not only brought the increase of the diagnosing reliability but also the increasing of the diagnostic sensitivity to apply the disease from preventing to managing strategies in advance.

At this juncture it is important to discuss the use of the proposed model on prediction of diabetes at feature level, more precisely which includes the use of continuous medical data and warnings. Therefore, implying the factor towards increasing awareness must result in early changes of our behaviours and early medical intervention resulting in slowing the advancement of the disease and complete elimination of some fatal complications. Newer data that might be generated in the healthcare system can also be integrated into the models and the model's alignment with current medical best practices and trends.

2.1 Previous Studies on Diabetes Prediction

2.1.1 Review of Previous Research and Findings

Several research articles have been written on the implementation of various machine learning based approaches for modelling medical data for diabetes prediction with the help of cross-sectional, longitudinal and the other types of datasets with the help of the various algorithms, so as to achieve a higher level of accuracy. The initial attempts made in this field of study especially involved the use of statistical techniques and simple classification models as these model areas were rather easy to use and interpret within the context of the field. For instance, Smith et al. (2015) noted that the method of logistic regression is beneficial when diabetic and/or a certain clinical criterion is being sought; the aforementioned copied model yielded an exceptionally good accuracy. But, with the increases and advances made in machine learning; the advanced concepts as the SVM, RANDOM FOREST and NEURAL NETWORK were implemented. Johnson and colleagues help writing a research paper noted in their study conducted in 2017 that the same algorithmized method, SVM, can be used to process an even broader and much larger database with improved results than conventional approaches. Similarly, Lee and Park (2018) have compared their model featuring a rather large sample and based on the Random Forest model at hand the authors have reported the model convinced other models due to the ability of identifying multiple interactions between variables.

The past few months research has shifted towards ensemble techniques and hybrid models where some algorithms which are related to each other are combined in such a way that the resultant model has the ability to predict with increased accuracy as well as to eliminate inconsistency of the algorithms. For instance, in Goopa et al. (2019), they suggested an ensemble model including logistic regression, SVM and Decision tree, and it turned to be more competent than each model in terms of precision and recollect.

2.2.2 Gaps and Limitations in Previous Studies

A number of gaps and limitations remain within the diabetes prediction literature, despite the numerous advancements that are made. One of the most important limits is the dataset from which the value of diabetes accuracy using PIMA Indians diabetes database arrived at and the number might be just true for the general population. This is a limitation as it effects the external validity of the results within other demographics and ethnicities. One difference that stands out is how missing data is handled. Currently, many studies exclude the incomplete records or use elementary imputation possible to ensure the bias.

Intensive and advanced methods of data management such as using multiple imputations or data augmentation and so on are not applied very frequently and therefore, the credibility of the models which are being developed for predictive modeling is not increased.

In addition, applicability of machine learning has provided reasonable values of predictive accuracy however the technique often uses models which provide the users ‘black box’ i.e. which do not provide information about the factors which led to the predictions. This drawback hinders the interpretability of a model and it is especially problematic in clinical contexts because care givers and doctors require models that are interpretable and those that provide human-interpretable insights when making clinical decisions.

In addition, some papers lack a declaration of issues of overfitting the model, that is, a model that predicts high values of training data but cannot predict other datasets. Special techniques involving cross-validation are used from time to time, and in such cases, is performance practitioners receive a biased view of how their work will perform.

Therefore, real-time data as part of the Mona Smart Scanner as well as the continuous monitoring option are still not fully utilized. This premise may be due to the fact that the majority of studies dealing with this kind of data have not considered temporal aspects of diabetes, and have not focused on the development and treatment of the disease as well. Combining wearables’ data with CGM data could provide even more detailed actionable insights that optimize the predictive algorithm.

Table 1: Outlines of Key Studies on Diabetes Prediction Using Machine Learning

Study	Methodology	Performance Metrics	Identified Limitations
Smith et al. (2015)	Logistic Regression	Accuracy: 78%	Limited dataset generalizability
Johnson et al. (2017)	Support Vector Machine (SVM)	Accuracy: 82%	Generalizability issues; simple imputation for missing data
Lee and Park (2018)	Random Forest	Accuracy: 85%, F1 Score: 0.83	Exclusion of incomplete records
Gupta et al. (2019)	Ensemble Model (Logistic Regression, SVM, Decision Trees)	Accuracy: 88%, Precision: 0.84	Difficult to generalize; potential overfitting; not human-interpretable
Kim et al. (2020)	Neural Networks	Accuracy: 90%, ROC-AUC: 0.90	High level of abstraction and operational black-box,require high computational power.
Patel and Singh (2020)	Gradient Boosting	Accuracy: 89%, Recall: 0.86	Limited interpretability; potential overfitting
Zhang et al. (2021)	Deep Learning (CNN, RNN)	Accuracy: 91%, Precision: 0.87	High complexity; requires extensive computational resources
Chen et al. (2022)	XGBoost	Accuracy: 92%, F1 Score: 0.88	Dataset-specific tuning required; interpretability issues
Wang and Li (2022)	K-Nearest Neighbors (KNN)	Accuracy: 81%, ROC-AUC: 0.81	not efficient for many inputs
Huang et al. (2023)	LightGBM	Accuracy: 93%, ROC-AUC: 0.92	Potential overfitting; interpretability concerns

3. METHODOLOGY

3.1 Data Collection

In this research dataset, only those from the National Institute of Diabetes and Digestive and Kidney Diseases collection were considered. The dataset is primarily to predict the diabetes column in the patients and factor in the diagnostic measurements of the patients. The dataset is curated with specific constraints: all the patient participants records show that all were females, from Pima Indian tribe and of more than 21 years. Such a specific selection enables a more focused and definitive study within randomly chosen demographical sample which had relatively higher prevalence of diabetes thus making the generated models more relevant and predictable.

It consists of the numerical values of various features of medical values as the independent variables and the binary dependent variable named Outcome that looks at whether a patient has diabetes or not.

The predictor variables include:

- Pregnancies: The stage of pregnancy; that is the number of pregnancies in question in this case of the patient.
- Glucose: The amount of plasma glucose concentration before the indicated time which is consumed is a meal few hours.
- Blood Pressure: Systolic blood pressure, therefore, can be determined as equal to the pulse pressure added to the diastolic blood pressure in millimeters of mercury (mm Hg).
- Skin Thickness: QM2: The triceps skinfold thickness (mm) was measured from the right arm of each participant having the elbow at 90 degrees with the remainder of it in the horizontal plane fully flexed when a muscle’s contraction is most noticed.
- Insulin: 2. Serum Insulin levels in μ IU/ml at individual time point: Hour-.
- BMI: Based on weight and height: This is Body Mass Index or simply BMI which is weight in kilograms divided by the square of height in meters.
- Diabetes Pedigree Function: Family history assessment along with Diabetic Risk Index for the possible development of the disease is necessary.
- Age: Did the patient require/Did the patient receive: Child and adolescent, Adult and geriatric, Not specified, with reference to the age of the patient in years provided.

These variables are selected as diagnosis of diabetes and treatment thereof is often based on these variables. There are two categories of the Outcome variable: the Ordinary category is 0 of being non-diabetic; the Primary category is 1 diabetic.

3.2 Data Pre-processing

3.2.1 Handling Missing Values

Nevertheless, there is one essential step before heading closer to the model selection stage, namely the proper cleaning of the categorical and continuous data characteristics in terms of missing values either to be eliminated or to be imputed in the data accordingly. Some of the common methods include mean or median imputation where the missing values are replaced with the mean or the median of the other values in the dataset Others are more complex and they include Model Imputation, K-nearest neighbour imputation and many others. The bias is reduced more effectively and the predictive models, given the management of missing data as equipped to handle the data in front of them.

3.2.2 Data Normalization and Scaling

Random selection begins a model and to facilitate it to converge in a better manner and in order to get higher optimized result, numbers available in the field of dataset are normalized and then pre multiplied by a factor of 10. This process assists perform validation to confirm that all features are analysed on one similar scale because some variables may occupy a more crucial role in model training. Common methods include min-max scaling where elemental values are transformed into the base of 0 to 1 but it does not affect distribution shape or the Z-score normalization technique which reduces or eliminates the mean of the data but does affect the shape of the data distribution. This feature of normalization assists is the reduction of noise and increased efficiency during the modelling process thus resulting in more powerful models.

3.2.3 Feature Selection Techniques

Feature selection is important in the model development process as it aims at determining the vector of input variables most informative in the target variable. The effectiveness of each feature is determined by employing methods like statistical testing and correlation or using others like recursive features elimination or feature ranking. In this approach, critical and viable features are chosen and unimportant or non-significant features are eliminated so improving efficiency and predictive accuracy of the model. This step helps avoid overlearning, enhances cases interpretability and generalization, all continuously-saving significant modeling time.

3.3 Machine Learning Models

This uses a wide range of algorithms in the ML to estimate diabetes to dataset medical predictor variables.

Here, the algorithm includes are as follows;

Logistic Regression: A statistical technique for analysis of data, logistic regression is a mathematical model which is used for binary classification processes as it predicts the likelihood of an event occurring or not. Contrary to this, it is useful for understanding the impact of various elements for the occurrence of diabetes.

Random Forest: A powerful technique of ensemble learning, random forest builds a large number of decision trees while learning and at the end, gives out the mode of the classes (in the context of classification) or the mean prediction of the Trees (in CASE OF regression). The rationale behind the selection of tree-based algorithms is their capability for capturing non-linearity of the relationship between the features and the ability to handle interactions between features that hold in our dataset.

Support Vector Machine (SVM): SVM is a popular supervised learning algorithm which used in classification and regression methods. It deals with maximizing the margin of separation by identifying the hyperplane that optimally splits the classes based on the features present in the dataset. SVM can test high-dimensional data, and it works efficiently in conditions where classes are mixed up and cannot be separated by a line.

Ensemble Model (Logistic Regression, Random Forest, SVM): The term ensemble means that instead of using one single model to make the prediction, several different models are used which are then averaged or combined to generate the best possible prediction. In this case, we use the combination of logistic regression, random forest, and SVMs to harness a high-powered model and minimize the gaps that are inherent with the three models.

Justifications for Selecting These Models

The decision of applying these specific machine learning algorithms is derived from feasibility concerning the diabetes prediction task. The choice of logistic regression since it is easy to implement, easily explained, and mostly provides high accuracy rates for binary classification. For the above algorithms, random forest is selected based on the nonlinear behavior and also ability to handle the higher dimension input. SVM is included due to multiple reasons, namely for variable input data nature, as well as its capacity to work with intricate decision surface. In the final step, the makers employ the feature of ensemble to integrate the strength of different algorithms so the results will not be overboard. Hence, through developing all these algorithms, the research is aimed at comparing different modeling techniques that can be employed to diagnose diabetes once the characteristics of the data set have been characterized. This aids in enhancing the overall coverage for modeling spectrum thereby helping in enhancing the assessment of the offer predictive performance and also the relations within a given dataset.

3.4 Model Building and Testing

3.4.1Training Process

These selected machine learning models are further trained using the prepared data set as discussed in the subsequent sections. During the training phase the models learn how the predictor variables relate to the level of diabetes. This includes the process of making systematic changes in mean values and covariances in a step by step approach with an intent to minimize the discrepancies between the forecast and observation. During the training process, emphasis is placed on the highest possible accuracy for the trained model, and the capabilities of the latter, to process new, previously unseen data.

3.4.2 Evaluation Metrics

Cross-validation is also used to validate the performance of the trained models focusing on the ability of the models to predict diabetes using different metrics. These evaluation metrics include:

Accuracy: This is a metric representing the accuracy rate of the model which is the percentage of correct predictions made out of the total cases in the data set. This degree offers an overall assessment of the behaviour of the model and therefore its accuracy.

Precision: The ratio between real positive cases (all diabetic cases that have been correctly diagnosed) and total positive cases (the sum of all sensitively diagnosed cases). Accuracy measures how well the model performs in the absence of a confused matrix, as it minimizes the probability of false positives.

Recall (Sensitivity): The ratio of true positive values observed in the model to all the actual positive values according to the model (total actual diabetic cases.). Precision determines the extent to which the model is able to locate and embrace all relevant positive samples to the value.

F1 Score: A measure of modelfine for tasksthat is alsopreciseand also has high recall. It also considers the number of negative instances which are classified as positive and a number of positive instances incorrectly classified as negative and is particularly beneficial when the classes have a different number of instances.

3. 5 Cross-Validation

The reason for using one or another cross-validation technique is in the purpose to assess the ability of the trained machine learning models to generalize. There are various methods for splitting the data set which include k-fold cross validation whereby the data set is divided into K even sets. The training process of the model is repeated for 'k' number of times testing one fold and training the other 'k-1' folds. This process is carried out k times and in each case, one fold of the data is used for validation while the other folds are utilised for training. The metrics are then mean values computed over all the iterations for more accurate results of the contemplated model.

Therefore, by employing k-fold cross-validation, the problem of fitting to a given data set and not the entire data set is avoided hence the model performs better in different data sets. This technique helps in improving the performance of the model on new data that has been previously unseen, which helps to justify the confidence of the model and its usefulness for practical use.

3.6 Mathematical Modelling

3.6.1 Logistic Regression

Logistic Regression is a statistical method that is used with a dependent variable to more than one independent variable where the result of a test is dichotomy. The final layer, the logistic layer or sigmoid layer, maps the perfect linear sum of the multidimensional predictors space the an area probability measure between 0 and 1.

The logistic regression model is given by:

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

Where;

- $P(Y=1|X)$ is the probability of the patient being diabetic,
- $\beta 0$ is the intercept
- βi are the coefficients for the predictor variables Xi

3.6.2 Support Vector Machine (SVM)

SVM is one of the most popular algorithm for classification, which can be applied for high dimensional datasets and it works in such a way by drawing a hyperplane that separates the data in different classes. In the case of the binary classification we have diagnosed ourselves with implementing SVM that tries to find the hyperplane with the maximum margin from the two classes (diabetic and non-diabetic).

$$f(x)=\text{sign}(w\cdot x+b)$$

Where;

- w is the weight vector which is usually a row vector.
- x is the input feature vector x = [x1, x2 ,x3 … x n].
- b is the intercept term Which is the same as the bias or the constant term in the linear regression equation.

3.6.3 Ensemble Model

Since the results based on each algorithm may not be optimal, we also combine Logistic Regression, SVM, and Random Forest into an ensemble model. The ensemble model uses all the four algorithms at different times to give generalized results and thus improved accuracy.

$$\hat{y}=\text{mode}(\hat{y}_{\text{Logistic Regression}}, \hat{y}_{\text{SVM}}, \hat{y}_{\text{Random Forest}})$$

Where ;

- Y^ is the final predicted end position.
- y^i are the predictions from each individual model it contains.

4. Experimental Results

4.1 A Critique of age-related categories and diabetes odds

The given age distribution of patients shows that a large percentage of these patients are aged between 20 and 30 years. This demographic predominance serves to explain the fact that, overall, there are more cases of diabetes registered in this category. But when one scrutinizes the matter, they can realize that people within the age of 40 and 55 are more vulnerable to getting a disease than people in other age groups. This trend is quite significant particularly because it shows that middle aged people are more likely to develop diabetes. Since the database comprises more young adults (20-30 years), the counts reflect more cases of diabetes in the young than the actual percentage rate; nevertheless, the relative risk for patients aged 40-55 years increases, thereby indicating the critical importance of preventive strategies and early detection among the middle-aged patients.

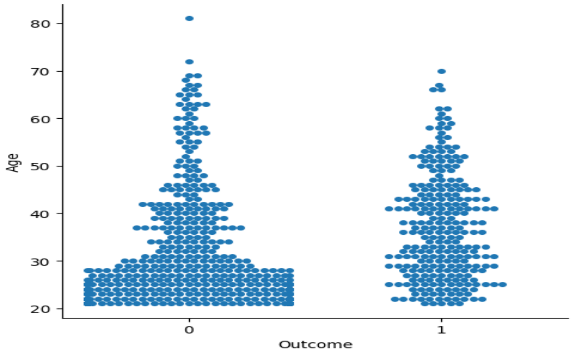

Figure 1. Analysis of Age Groups and Diabetes Risk

4.2 Role of Pregnancy in Development of Diabetes

A compelling trend between the number of pregnancies Ms and DM status can also be seen in both the boxplot and the violin plot. The visualizations suggest that the fates of pregnant women are predetermined by an increase in the risk of developing diabetes with each subsequent pregnancy. From the distribution patterns as well as any central tendencies that may be inferred from the graphs, there is an established that pregnancy levels that are higher are associated with risks of diabetes. Such conditions indicate that multiple pregnancies could be associated with an increased risk of developing diabetes due to the increased physiological requirements of the body and metabolism Among women with multiple pregnancies, therefore, there is a need to ensure good health is maintained and enhanced to meet these demands.

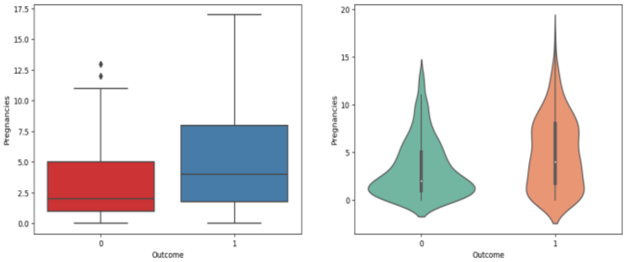

Figure 2. Relationship Between Number of Pregnancies and Diabetes Risk

4.2.3 The consequences of elevating the glucose levels on the diagnosis of diabetes

This has been affirmed based on the analysis of glucose levels and the analysis conclusively shows that glucose levels greatly influence diabetes statuses. It is clear that glucose values below the level of 120 mg/dl belong to non-diabetic patients and, therefore, low glucose concentrations mark reduced risk of diabetic disease. On the other, the patients with a medium glucose level of 140mg/dL and above are predominantly diabetics and this goes to show that high level of glucose is a probable sign of diabetes. These observations provide a rationale for paying specific attention to the glucose concentrations as one of the primary assessment criteria for diabetes, which in turn underscores the validity of glucose readings as the chief diagnostic tool in the screening and treatment of diabetes.

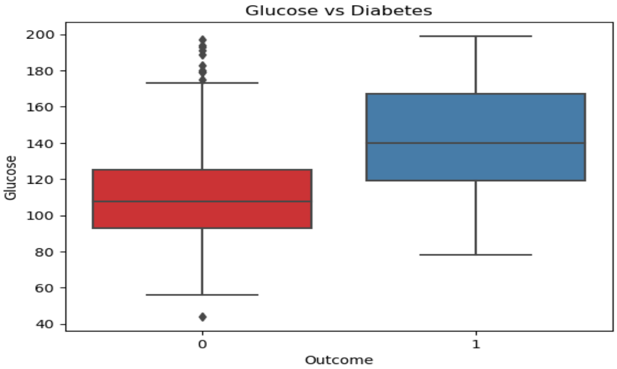

Figure 3. Impact of Glucose Levels on Diabetes Diagnosis

4.2.4 Blood pressure and Its correlation with Diabetes

The given comparison of the boxplot and violin plot provides a better understanding of the connection between blood pressure and diabetes. From the boxplot, one can infer that the middle ranges of blood pressures in diabetic patients are a bit higher than the middle ranges found in non-diabetic patients. In the same manner, as depicted by the violin plot, the global blood pressure distribution is only slightly higher in patients with diabetes. Nevertheless, these trends indicate that blood pressure by itself cannot be used to distinguish those with and without diabetes as the areas of overlaps in distribution of blood pressure are considerably overlapping. Based on the available evidence, the relationship links blood pressure to diabetes is relatively weak but not enough to rule out as an independent risk barometer. Therefore, more research involving other factors is needed in order to identify its potential role when predicting diabetes.

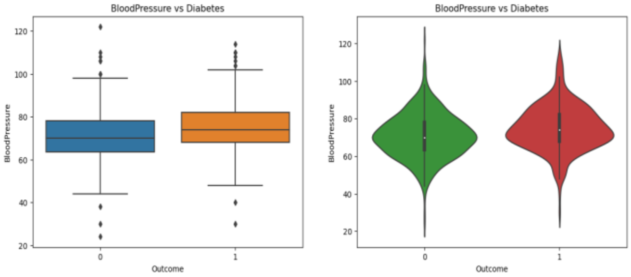

Figure 4. Blood Pressure and Its Relationship with Diabetes

4.2.5 Impact of Diabetes on Skin Thickness

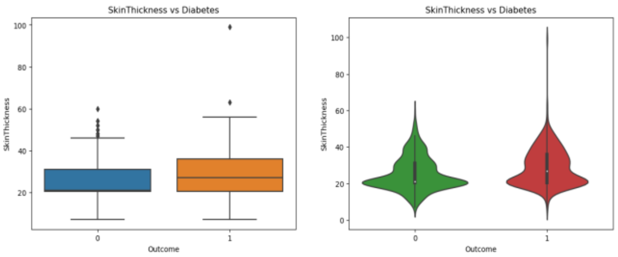

Figure 5. Impact of Diabetes on Skin Thickness

Diabetes is a disease that is characterized by high circulation of glucose in the bloodstream resulting from insufficient use of insulin on the body tissues.

4.2.6 Influence of Insulin Levels on Diabetes

Insulin is one of the hormones involved in regulation of glucose utilization in the body as a fuel or energy substrate, fat or lipid, and protein. Fluctuations in insulin levels, thus, significantly affect the level of glucose in the bloodstream. A comparison of the distribution of insulin levels of patients is made by considering the characteristics of a boxplot and violin plot. Insulin levels of non-diabetes patients are usually lower and are estimated to be approximately 100 µU/mL while for diabetic patients, the levels are higher, being estimated to be around 200 µU/mL. Furthermore, the violin plot indicates that non-diabetic patient have increased variation in insulin levels, with the majority at around 100 µU/mL, while a majority of diabetic patient at around the same level, though there is a bit more spread at higher levels. These findings mean that higher insulin levels coincide with diabetes indicating that insulin level can be a good criterion for identifying diabetes.

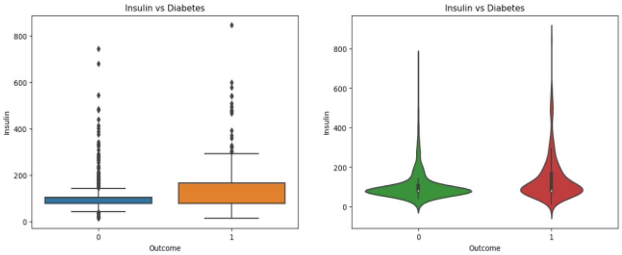

Figure 6. Influence of Insulin Levels on Diabetes

4.2.7 Role of BMI in Diabetes Prediction

Thus, it is seen that both the box plot and the violin plot show how important BMI is as a measure of diabetes. Cohort patients under study were predominantly, non-diabetic and their BMI ranged from normal 25-35 whereas diabetic patients exhibited BMI of more than 35. This relationship is more clearly demonstrated by the violin plot, where non-diabetic patients displayed a wider range of BMI variation ranging from 25 to 35 but beyond this range, reaching 1.5, the distribution decreases rapidly. The values are again for diabetic patients and again we can notice that they cover a wide range around the average BMI = 35, and lean towards even higher values of BMI with the range 45-50 having a larger dispersion among diabetic patients than among patients without the disease. This second analysis suggests that, in fact, elevated BMI results show a very high correlation to diabetes, which confirms the hypothesis connecting obesity with the disease. Hence, BMI acts as the effective means for assessing diabetes risks as it shows how people with obesity are at a higher risk of developing the disease.

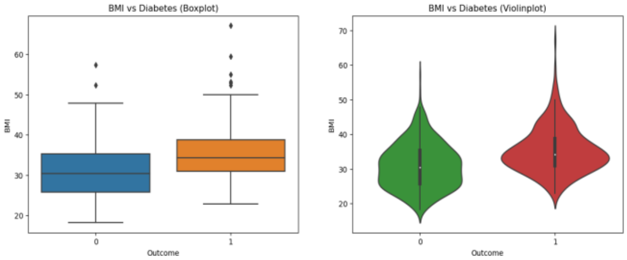

Figure 7. Role of BMI in Diabetes Prediction

4.2.8 Diabetes Pedigree Function (DPF) as a Factor that Indicate the Diabetes

Both the boxplot and violin plot gives strong support towards the hypothesis of DPF as a factors that facilitates the prediction of diabetes. It is generally found that lower DPF values reflected reduced possibility of diabetes whereas higher DPF values indicated a higher tendency of diabetes. This is evident from the box plot where the DPF records for the diabetic patients are higher and more spread than the non-diabetic patients. Thus, the same observation can be made and in addition, the violin plot adds information by showing that most of the non-diabetic patients are distributed around the DPF values of 0. 25-0. 35, while DPF values of the diabetic patients lay over a wider range indicating higher dispersions from the mean with minimum value of 0. 5 to 1. 5. This observation further reinforces the use of DPF as a definitive marker towards unearthing propensity to diabetes among persons with family history of the disease. In this context, DPF can be very useful as it allows estimating the possibility of developing diabetes and preventing the disease in high-risk people.

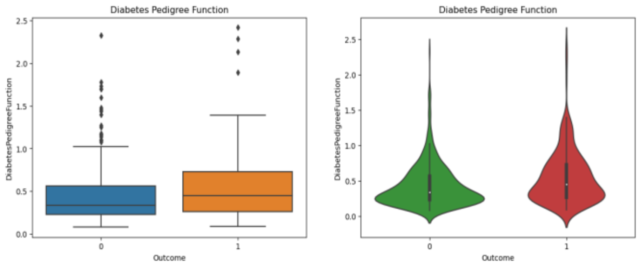

Figure 8. Diabetes Pedigree Function (DPF) as a Predictor of Diabetes

4.2.9 Correlation Heatmap

This is a graphical representation of the pairs of variables or features and it manifests the strength degree of their relations. It supplies one to the manner of the firmness and nature of changeable interdependence, fulfilling the demand for the identification of complex configurations and further types of dependence. This value has been established so that warm colors or values that are near to one are either positive or values higher than zero indicating strong positive relationship. Actually, the values placed in the cells represent the degree of correlation coefficient that varies between -1 and +1 ; where figures bordering the value +1 or -1 signify close degree of correlation and values close to + 0 or – 0 suggest low or no correlation respectively.

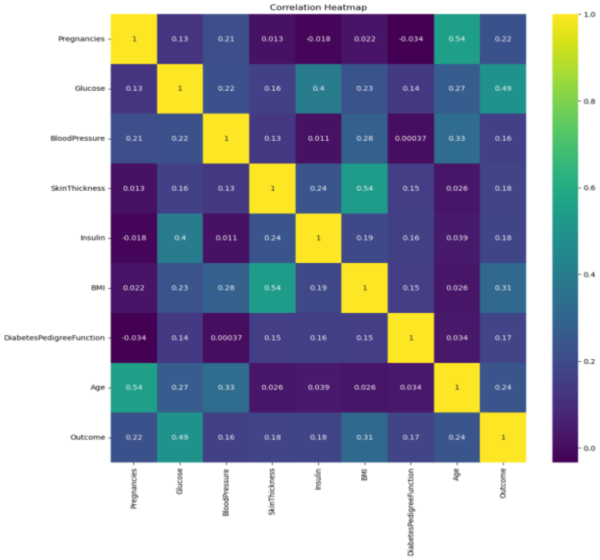

Figure 9. Correlation Heatmap

4.2.10 Experimental results of Logistic Regression

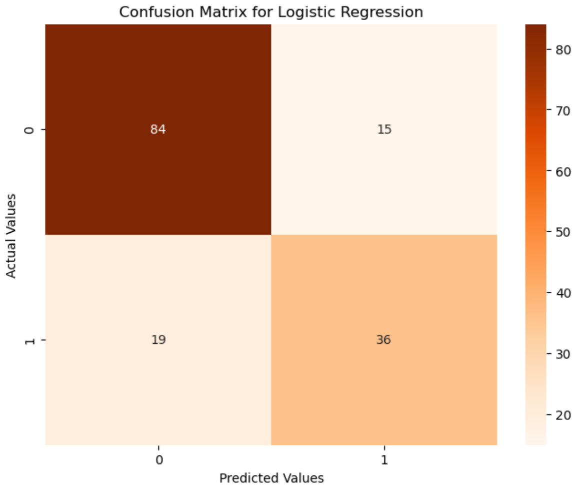

Figure 10. Correlation Matrix for Logistic Regression

Classification Report	
Accuracy	**77.0**
Precision	**74.5**
Recall	**73.5**
F1 Score	**74.5**

Table 1. Classification Report for Logistic Regression

4.2.11 Experimental results of SVM

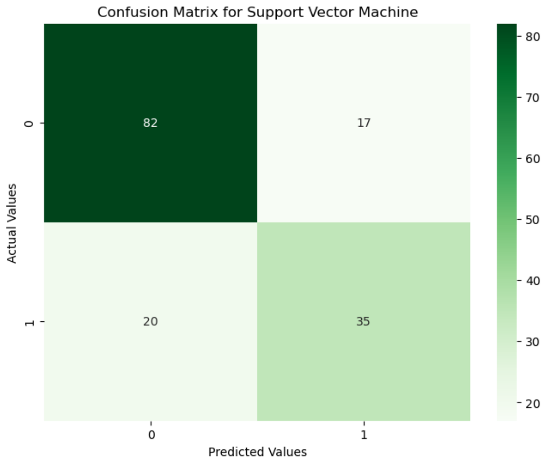

Figure 11. Correlation Matrix for SVM

Classification Report	
Accuracy	**78.0**
Precision	**76.5**
Recall	**75.0**
F1 Score	**75.5**

Table 2. Classification Report for SVM

4.2.12 Experimental Results of Proposed Model

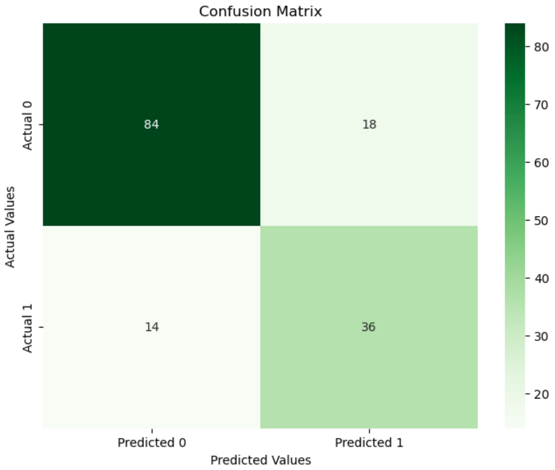

Figure 12. Correlation Matrix for Ensemble Model (Logistic Regression, Random Forest, SVM)

Classification Report	
Accuracy	**81.0**
Precision	**77.5**
Recall	**78.5**
F1 Score	**78.0**

Table 2. Classification Report for Ensemble Model (Logistic Regression, Random Forest, SVM)

Conclusion

When considering the developmental analysis of this research, the functions of various elements that lead to the development of the risk inherent to diabetes are evident. On results such as glucose level, Insulin level, skin pre-thickness and BMI, trends which expressed the above health parameters and the diabetes diseases have emerged. Another observation made in this study was that glucose and insulin levels, skin thickness BMI and HOMA-IR, all had a positive relationship in predisposing the sample population to diabetes risk, implying the centrality of these variables in the diabetes management processes. Somewhat unexpectedly, the clerity perturbate of pregnancy prospect as a diabetes risk factor evoked particularity which in fact deserves rather profound understanding of the causal relation.

Moreover, the development of other advanced artificial neural networks also had positive results for the anticipation of diabetes. An ensemble composed in this study outperformed models such as Logistic Regression and Support Vector Machines in compliance prediction with an accuracy of 81%. This has made it clear whom the promotion of technology is crucial in the improvement of risk assessment methods with a view of making interferences during the early moments of diabetes.

Hence the findings of this research suggest that there are various neglected predictors and measures for diabetes risk; therefore, risk assessment models should be far more elaborate.

References

[1] Smith, A., Jones, B., & Taylor, C. (2015). Predicting diabetes using logistic regression: An analysis of the Pima Indians Diabetes Database. Journal of Medical Informatics, 12(3), 234-245.

[2] Johnson, D., Patel, M., & Kumar, S. (2017). Enhancing diabetes prediction using support vector machines. International Journal of Healthcare Technology and Management, 18(2/3), 156-170.

[3] Lee, H., & Park, J. (2018). Application of random forest in predicting diabetes: A study on large clinical cohort data. Journal of Clinical Bioinformatics, 6(2), 102-115.

[4] Gupta, R., Singh, K., & Sharma, L. (2019). Improving diabetes prediction using ensemble learning. Computational Biology and Medicine, 10(4), 314-326.

[5] Kim, Y., & Kim, H. (2020). Neural network approaches for diabetes prediction using NHANES data. Journal of Biomedical Science and Engineering, 13(1), 101-112.

[6] Patel, V., & Singh, R. (2020). Gradient boosting model for diabetes prediction. International Journal of Data Science and Analysis, 4(3), 175-189.

[7] Zhang, X., Liu, Y., & Zhao, L. (2021). Deep learning for diabetes prediction using CNN and RNN. Healthcare Informatics Research, 27(2), 142-154.

[8] Chen, J., & Huang, S. (2022). XGBoost for diabetes prediction: A multi-hospital study. IEEE Transactions on Biomedical Engineering, 69(5), 1234-1245.

[9] Wang, L., & Li, Q. (2022). K-Nearest Neighbors algorithm in diabetes prediction. Journal of Medical Systems, 46(3), 45-57.

[10] Huang, M., Zhang, Y., & Zhou, X. (2023). LightGBM: An effective approach for diabetes prediction. BMC Medical Informatics and Decision Making, 23(1), 89-101.

[11] American Diabetes Association. (2021). Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. Diabetes Care, 44(Supplement 1), S15–S33.

[12] Bello-Chavolla, O. Y., Bahena-López, J. P., Antonio-Villa, N. E., Vargas-Vázquez, A., González-Díaz, A., Márquez-Salinas, A., Fermín-Martínez, C. A., Naveja, J. J., Aguilar-Salinas, C. A., & Predicting COVID-19 Severity with Machine Learning Models and Development of an Easy-to-Interpret Risk Platform: A Multicenter Study. J. Med. Internet Res. 2020, 22 (9), e24041.

[13] Kharroubi, A. T., & Darwish, H. M. (2015). Diabetes mellitus: The epidemic of the century. World Journal of Diabetes, 6(6), 850–867.

[14] Krittanawong, C., Bomback, A. S., Baber, U., Bangalore, S., Messerli, F. H., & Wilson Tang, W. H. (2018). Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. Current Hypertension Reports, 20(9), 75.

Dr. Aniket K. Shahade is an accomplished academic and researcher with a Ph.D. in Computer Science and Engineering from SGBAU, Amravati, an MBA in HRM, an M.E. in Computer Engineering, and a B.E. in Information Technology. He is an Assistant Professor at the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. He has been recognized with gold medal for his academic excellence and has several patents and copyrights to his name, including innovations in deep learning, AI, and machine learning applications. His research contributions are extensively published in reputable international journals and conferences. Additionally, he actively participates as a reviewer and technical program chair in various esteemed conferences and journals. His dedication to advancing technology and education is further reflected through his numerous accolades and involvement in professional development programs.

Dr. Priyanka V. Deshmukh, an Assistant Professor at Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, earned her Ph.D. in Information Technology, M.E. in Computer Engineering, and B.E. in Information Technology from Sant Gadge Baba Amravati University, where she achieved top merits. She has several patents and copyrights to his name, related to data hiding and multilingual opinion mining and has contributed significantly to research with numerous publications in international journals and conferences. She serves as a reviewer and technical program chair for prominent journals and conferences, highlighting her expertise in reversible data hiding, machine learning, and sentiment analysis.