International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Diabetes Prediction and Apprehension with Focus Both on Clinical and Non-Clinical Factors

Aditya Gupta[1], Angad Singh[2], Maharshi Jani[3], Yuvraj Salaria[4], Dr. Vani Hiremani[5], Dr. Sudhanshu Gonge[6*], Dr. Ketan Kotecha[7]

Submitted: 28/08/2023 **Revised**: 23/10/2023 **Accepted**: 03/11/2023

Abstract: An industrial revolution has changed the daily lifestyle of engineers, doctors and common people. This is due to a lot of experimental work performed and carried out in the food industry, IT sector, agriculture industry, automobile industry, etc. It has an impact on the diet of the stakeholders. Due to this the enzyme generation gets reduced and may lead to low production of insulin. As insulin is one of the important parts of the blood which controls all properties of plasma, water, enzymes, protein, vitamins and minerals, this causes diabetes to the patient. It is the essentially the most common and widespread chronic disease in the world. In our research paper both clinical and non-clinical parameters are considered such as Insulin, Glucose, BMI, smoking, stress, BP, Junk food etc. From this the aim of the research is to emphasize on both kinds of factors that affect a person's probability of Diabetes detection. There are several techniques such as models of SVM, Decision Tree, Random Forest and Logistic Regression which were found useful for predicting and apprehending the features to identify diabetes. We have done comparative analysis of techniques to observe the output after applying clinical and non-clinical factors.

Keywords: *Classification, Regression, Probability, Machine Learning, Statistics.*

1. Introduction

The era defined by unprecedented technological advancements, the integration of machine learning and healthcare has come out as a pivotal force in the quest. It is used to improve patient care and enhance medical decision-making. Engineers, physicians, and regular people all live different lives today as a result of the industrial revolution. This is because there has been a great deal of experimental work done in the food business, IT sector, agricultural industry, automobile industry, etc. It influences how the stakeholders eat. It has become a challenge to society and is difficult to cure. It can be prevented by taking the major step when the percentage of insulin is less at initial level. There is the need for a blood smear device and a combination of machine learning algorithms which helps for prognostication and providing awareness to the patient before increasing the sugar level. Its multifaceted nature of the patient's blood, characterized by varying risk factors, complex etiology, and a diverse range of clinical manifestations, necessitates innovative approaches for early detection and effective management.

Machine learning, with its ability to discern intricate patterns within vast datasets, has ushered in the daily lifestyle of predictive healthcare. By harnessing the power of data, machine learning models have demonstrated remarkable potential in predicting the onset and management of diabetes. This report embarks on an exploration of the intersection between machine learning and diabetes, with a particular focus on the predictive capabilities that these computational tools offer.

The aim of this research is to conduct a comprehensive analysis of the methodologies, algorithms, and models utilized in the domain of 'Diabetes Prediction Using Machine Learning Algorithms.' Additionally, this study aims to explore the progress made in this field, identify the challenges faced, and discuss the potential for future advancements. Furthermore, it will also explain the advantages after predicting the target by selecting the features of blood and helps to maintain a sensitive health record of the diabetic patient. This will also reduce the doctors effort and help in advancing medical science.

[1] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India aditya.gupta.btech2020@sitpune.edu.in*
[2] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India angad.singh.btech2020@sitpune.edu.in*
[3] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India maharshi.jani.btech2020@sitpune.edu.in*
[4] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India yuvraj.salaria.btech2020@sitpune.edu.in*
[5] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India vani.hiremani@sitpune.edu.in*
[6] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India sudhanshu.gonge@sitpune.edu.in*
[7] *Department of Computer Science and engineering, Symbiosis Institute of Technology, Symbiosis International University (Deemed University) Pune, India director@sitpune.edu.in*
** Corresponding Author Email: sudhanshu.gonge@sitpune.edu.in*

2. Literature Survey

Amelec Viloria (2020) et al. mention that Diabetes mellitus is a disease which cannot be easily diagnosed as several factors play a pivotal role and a simple blood test cannot provide enough information to give an ultimatum that is trustworthy. The analysis of factors that affect the diagnosis are subject to human error hence we need to make use of the Machine Learning model which once trained, can be relied upon. Diabetes is caused by genetic factors; mutant genes in the chromosome 6 (responds to antigens).[1]

In the study conducted by Viloria et al., Support Vector Machine (SVM) was employed to predict the medical diagnosis of Diabetes. The researchers used Age, BMI (Body Mass Index), and blood glucose concentration as input indicators for the SVM model. These parameters were utilized to predict the diagnosis, which was classified as either positive or negative for Diabetes. To evaluate the performance of the SVM model, a 10-fold cross-validation method was employed. This validation technique includes metrics such as accuracy, sensitivity, specificity, positive and negative prediction values, and confusion matrix to assess the effectiveness of the model in making accurate predictions. An accuracy of 95.36% was achieved as per dataset of patients from Columbia using this model. While the accuracy is 95.36% with the data that the model is trained on and also patients with same ethnicity but when it is used with patients with different ethnicity it goes down to 66.25%.[1]

In her paper, KM Rani (2020) discusses the fundamental distinction between Type 1 and Type 2 diabetes. In Type 1 diabetes, the immune system is compromised, leading to a failure of cells to produce sufficient insulin. On the other hand, Type 2 diabetes occurs when body's cells produce insulin, but in lower quantities. Type 2 diabetes is the most prevalent form, accounting for approximately 90% of all diabetic cases. Additionally, Rani mentions Type 3 diabetes, which is gestational diabetes. This type occurs due to an increase in blood sugar levels in pregnant women, presenting a temporary form of diabetes during pregnancy. The accuracy of Random Forest classifier comes out to be highest i.e. 97%. Also, there might be a case of overfitting because of extreme accuracy in the Decision tree model.[2]

Approximately 2 to 5 million people lose their lives because of diabetes each year and this can be decreased by early diagnosis and treatment. According to Aishwarya Majumdar et al. (2019) current human lifestyle which is filled with consumption of unhealthy products and an inactive and lazy attitude is what has been one of the main reasons for the growth of such diseases. Predictive analysis is the method to go forward with as it provides accurate results and optimizes resources and saves time. K means clustering is effective to form clusters of two important attributes which highly influence the whole data like Glucose and BP.[3]

The paper done by Nongyao Nai-arun (2015) et al. states that almost half of the diabetes in South East Asia is undiagnosed and hence needs attention of the concerned authorities as they need to spread awareness of the disease. Also from [4] we learned that there are about 12 deaths out of 100 thousand people diagnosed with diabetes. Bagging and boosting techniques which can be applied to increase robustness of model after implementation of models.[4]

Kakoly (2023) et. al. with their paper explaining dietary habits of people in Southeast-Asian countries also add different influential factors other than those in datasets from developed countries. They studied using both clinical factors such as BMI, pregnancies etc and dietary factors like eating vegetables, fish etc. These habits keep on changing in a developing nation which has good economic growth. Principal Component Analysis (PCA) is used in [5] that eliminates the features and observed that most information falls where variation is greatest. Also Information gain was used in [5] which helps choose important features according to how relevant it is to a class. Various groups of features were also created that compared accuracy results for clinical and non-clinical factors. More clinical factors need to be included and use of ensemble methods is advisable in such an experiment.[5]

In their (2023) study, Osama R et al. utilized a deep learning technique, specifically the Deep Belief Network, to categorize and predict the development of diabetes. The research involved a structured approach that encompassed data collection, pre-training, and classification processes for forecasting. By employing this deep learning methodology, the researchers aimed to enhance the accuracy and efficiency of diabetes prediction, providing valuable insights into the disease's progression. The DBN process has proved to be of higher accuracy. Also, deep learning goes beyond machine learning and data mining, it helps in finding additional illness.[6]

Yahyaoui (2019) et. al. in their study compares three different learning-based methods for predicting diabetes using a specific dataset. The main objective is to evaluate the efficiency of both conventional machine learning and deep learning methods in predicting diabetes. The utilized techniques comprise Support Vector Machines (SVM) and Random Forest (RF) within traditional machine learning, along with Convolutional Neural Networks (CNN) in the realm of deep learning. Existing research indicates SVM and RF's efficacy across diverse classification tasks, while CNNs have gained prominence for classification and recognition in recent times. These algorithms were selected to gauge their accuracy in diabetes detection, with the ultimate goal of creating a decision support system.[7]

E. Ismail (2023) et. al. with their research tells that diabetes is a serious chronic illness associated with severe complications like blindness, kidney disease, and heart

attacks. Improving technology for patient monitoring and early intervention is crucial. In this research, machine learning algorithms were applied to forecast diabetes utilizing a dataset comprising 16,698 individuals. Three specific algorithms, namely K-nearest neighbor, Decision Tree and Naïve Bayes were utilized with a focus on accuracy, sensitivity, and complexity. The outcomes revealed that both K-nearest neighbor and Decision Tree achieved 100% sensitivity and exhibited high accuracy rates (99.64% and 99.61%, respectively). Furthermore, all models demonstrated favorable time complexity, indicating their potential in enhancing early diagnosis and treatment approaches..[8]

M. A. R. Refat (2021) et. al. discussed in this paper that diabetes as a disease characterized by issues with blood sugar regulation arises primarily due to inadequate insulin production or the body's ineffective utilization of insulin. The crucial role of insulin lies in facilitating glucose transportation for energy, underscoring the consequences of unregulated diabetes, such as hyperglycemia and associated health issues. Referencing a specific paper [9], it outlines an experiment where multiple Machine Learning (ML) and Deep Learning (DL) techniques were compared for early diabetes prediction, using a dataset encompassing 17 attributes. The findings indicated that the XGBoost classifier outperformed other methods, achieving an exceptional accuracy level of approximately 100.0%. Additionally, alternative algorithms also exhibited significant accuracy rates, surpassing 90.0%. [9]

Varun Jaiswal (2021) et. al. in this paper[10] presented that diabetes, a metabolic disorder marked by prolonged elevated blood glucose levels, can lead to severe complications like ketoacidosis, cardiovascular diseases, and renal failure. Its global prevalence has surged, rendering it a significant health challenge. Timely diagnosis holds immense importance for effective treatment and complication prevention. To identify patterns for diagnosis and treatment, machine learning algorithms and data mining techniques, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Naive Bayes, Partial Least Squares-Discriminant Analysis (PLS-DA), and deep learning, have been applied. However, current methods suffer from limitations, lacking cross-validation on diverse datasets, impeding practical application. This report synthesizes existing research on machine learning and data mining in diabetes prediction, with the goal of refining disease understanding, enhancing prediction accuracy, and ultimately improving treatment approaches, thereby reducing diabetes-related complications.[10]

Some of the shortcomings we noted are that existing methods for assessing diabetes diagnosis have not been tested on different data or on populations from different countries and are limited to the use of an approximate method.

Oza A. (2022) et. al. in this paper delves into the potential of data science methodologies to illuminate crucial subjects across scientific domains. Within the expansive realm of data science, machine learning stands out as a cutting-edge discipline devoted to enabling machines to learn from experience. Scholars in the field have conceived and implemented diverse data processing techniques with the aim of effectively categorizing and predicting symptoms in medical data. This study employs well-established predictive methods, specifically K-nearest neighbor (KNN) and logistic regression. The introduction of a predictive model aims to enhance and evaluate performance and accuracy by conducting a comparative analysis of the selected machine learning techniques. Notably, the paper does not explicitly disclose the method, number, and type of features chosen by the authors. [11]

F. M. Okikiola (2023) et. al. introduced a diabetes classification model that relies on a decision tree and a naive Bayes classifier. This model is structured into five key modules: document query, data classification, feature selection, and ontology development. The data collection module utilized the well-established PIDD for diabetes prediction. The feature selection module involves a series of steps to identify the most crucial elements within the customized diabetic dataset. This is achieved through a multi-step feature selection method. Initially, a forward selection strategy is employed, commencing with an empty set and gradually adding features to attain the highest classifier accuracy. Following this, the second stage employs backward elimination, systematically removing features that contribute the least to the decline in classifier accuracy from the entire set. This comprehensive approach is designed to optimize the accuracy of the classifier in diabetes classification.

Here Ratna Patil (2022) et al. employed the Mayfly Optimization Algorithm for the crucial task of feature selection, providing a distinct advantage in implementing the SVM classifier. Despite considering various machine learning models, such as random forest, regularized generalized linear model, and extreme gradient boosting techniques, there was no discernible clinical improvement in performance. The Mayfly Algorithm stands out for its balanced exploration of the search space and exploitation capabilities. It asserts that SVM, with the use of kernel functions, emerges as the optimal supervised classifier, effectively mitigating issues of overfitting and underfitting. Ratna et al. utilized a confusion matrix for both the training set and real-time dataset, showcasing the efficacy of the Mayfly-SVM. This approach proved to be unique and suitable, standing out as the best among all the comparisons made in their study. [13]

Han wu (2017) et al briefs us about Diabetes Mellitus, or DM, is a common ailment characterized by elevated blood glucose levels. It comes in two types: Type 1 Diabetes (T1DM), where the pancreas falters in regulating blood glucose due to damaged β-cells, and Type 2 Diabetes (T2DM), also known as Non-insulin dependent DM, marked by insulin resistance and deficiency. In addressing this challenge, our fellow turned to the realms of AI, Statistics, Machine Learning, and Database systems. Recognizing the need to establish predictive standards for the population more susceptible to DM, Dr. Saini and The Indian Weighted Diabetic Risk Score (IWDS) was proposed by Chandrakar. Han and Luo, meantime, proposed the Pair Wise and Size Constrained K-means (PSCK) approach to recognize high-risk people. K-means and Logistic Regression were just two of the models used. Evaluations such as K-fold cross-validation, confusion matrix, Kappa Statistic, and ROC curve were carried out to examine a model's performance. The researchers' model outperformed the competition with an accuracy rate of 95.42%. It is safe to claim that the model performed exceptionally well, delivering higher consistency and accuracy through improved K-means and Logistic Regression. In essence, this paradigm aids people in better understanding their health issues and directs them toward making healthier lifestyle choices. [14]

Diabetes stands as the most prevalent and arguably the most lethal ailment globally. Identifying it early not only arrests its progression but also aids in discerning its specific type. The researchers undertook a transformative approach by framing the task as a classification problem. Their model primarily relies on the concealed layers of a deep neural network, incorporating dropout regularization to prevent overfitting. Utilizing parameters like the binary cross-entropy loss function, the deep neural network achieved a predictive accuracy of 94.02174%. When applied to the PIMA Indian diabetes dataset, this accuracy rose to an impressive 99.4112%, outperforming existing models. To enhance the model's expressive capabilities, the deep neural network employs activation functions such as Sigmoid, Softmax, tanh, ReLU, softplus, among others. The resulting model, named DLPD, not only predicts the presence of diabetes but also forecasts the potential type of the ailment in the future, distinguishing between T1D and T2D. [15]

3. Methodology

The following diagram(Fig 1) is the methodology which was followed and the initial phase involved a meticulous data collection process, encompassing the acquisition of patient demographics, medical records, and lifestyle factors. Subsequently, the data underwent preprocessing, which entailed handling missing values, encoding categorical variables, and scaling features. Notably, categorical variables were cleaned and converted into boolean form.

To identify the most influential predictors, a feature selection process was implemented. Following this, the dataset was split into training and testing subsets, with an 80% and 20% division, respectively. Leveraging the training data, various machine learning algorithms, including logistic regression, decision trees, random forest, and support vector machines, were trained and evaluated using pertinent metrics such as confusion matrix, accuracy, precision, and recall.

The performance of the models underwent refinement through the process of model selection. The chosen model was then implemented for making predictions about the real world, with ongoing evaluation and adjustments as new data became available. This systematic approach ensures that the diabetes prediction model is both precise and understandable, empowering medical professionals to identify patients earlier and intervene effectively.

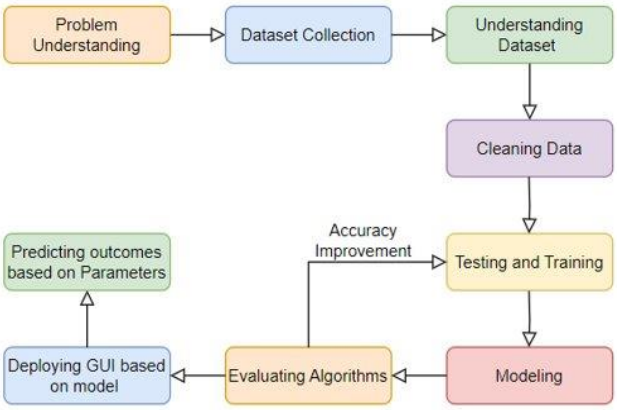

Fig 1: Proposed Methodology

3.1. Dataset

The template We have seen two major datasets where we have done both preprocessing and cleaning, so the PIMA dataset showed more promise in terms of our research as it contains clinical factors. The dataset has the following parameters: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI and DiabetesPedigreeFunction(family history). All these focus on the clinical factors of the problem and give a more understandable reason for direct impact on being diabetic or not. The dataset was numerical with few categorical parameters that needed our attention.

The other dataset which we saw contained various non clinical factors as well hence we also had to consider using it for concluding the effect of such factors on whether a patient is diabetic or not. This dataset has the following parameters: Family_diabetic, PhysicallyActive, smoking, alcohol, BMI, Sleep, SoundSleep, JunkFood, Stress, Pregnancies and more. It is more inclined towards the non-clinical factors hence it gives an indirect impact on diabetes. More so, these habits lead to the change in clinical factors.

Eating more junk food and not exercising can increase the chances of being diabetic as the blood sugar level can turn inappropriate for people in the age of 40 onwards.

3.2. Data Preprocessing

a) Information Gain: It quantifies the improvement in classification accuracy by measuring the reduction in entropy before and after splitting the data based on a particular attribute. Higher information gain values signify attributes that contribute significantly to the predictive power of the model.

Table 1: Information Gain

Pregnancies	0.688
Glucose	0.881
BloodPressure	0.821
SkinThickness	0.859
Insulin	0.459
BMI	0.880
DiabetesPedigreeFunction	0.891
Age	0.855

From our dataset we observed that Glucose and DiabetesPedigreeFunction produce a higher information gain which led to us selecting it in our model ahead and thereby improved our chance of getting a better accuracy and overall result.

b) Entropy: By strategically selecting attributes that minimized entropy through data partitioning, we aimed to enhance the model's ability to classify and predict diabetes outcomes accurately.

Table 2: Entropy

Pregnancies	3.401
Glucose	6.634
BloodPressure	4.597
SkinThickness	4.680
Insulin	4.674
BMI	4.660
DiabetesPedigreeFunction	0.242
Age	4.890

It has helped in forming the root as well as nodes of the decision tree and given initial understanding of what can be used for various models.

3.3. Visualization

Here we have done lots of methods that could not be added in the contents of this paper but we found these to be most appropriate ones which are mentioned and explained below. From these visualizations we have shown the understanding of the dataset without displaying it. It has also given how attributes are varying with other attributes. And well related they are to overall change.

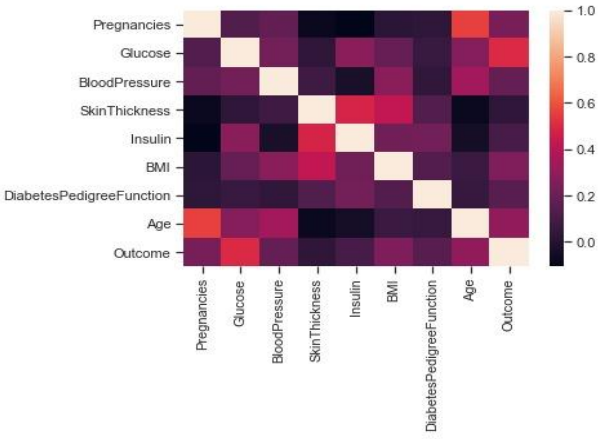

Fig 2: Correlation Heat map

From Fig 2 we were able to see that Age and Pregnancies, Insulin and SkinThickness give a high correlation and we made these four parameters paired so that they reflect well in the model's performance metrics.

Table 3: Variance of attributes

Pregnancies	10.119
Glucose	849.435
BloodPressure	120.592
SkinThickness	231.003
Insulin	6811.791
BMI	41.668
DiabetesPedigreeFunction	0.066
Age	113.500

Table 4: Covariance

	Pregnancies	Glucose	BP	SkTh	Insulin	BMI	DBF	Age
Pregnancies	10.119	12.011	6.273	-3.507	-27.656	0.482	0.025	19.093
Glucose	12.011	849.435	69.985	14.803	667.772	35.302	0.451	83.106
BP	6.273	69.985	120.592	12.756	-28.787	19.737	0.113	39.708
SkTh	-3.507	14.803	12.756	231.003	605.186	41.732	0.514	-12.087
Insulin	-27.656	667.772	-28.787	605.186	6811.791	112.755	4.652	-38.402
BMI	0.482	35.302	19.737	41.372	112.755	41.668	0.223	4.479
DBF	0.025	0.451	0.113	0.514	4.652	0.223	0.066	0.158
Age	19.093	83.106	39.708	-12.087	-38.402	4.479	0.158	113.5

SkTh ~ Skin thickness, DBF ~ DiabetesPedigreeFunction

From Table 3 and Table 4 we have produced the variance and covariance between the attributes, this allows us to be careful while using these in our models as there inverse variation can have irregularities while prediction. Also high variance of Insulin is dealt with caution as a wider spread of values leads to loss of predictability.

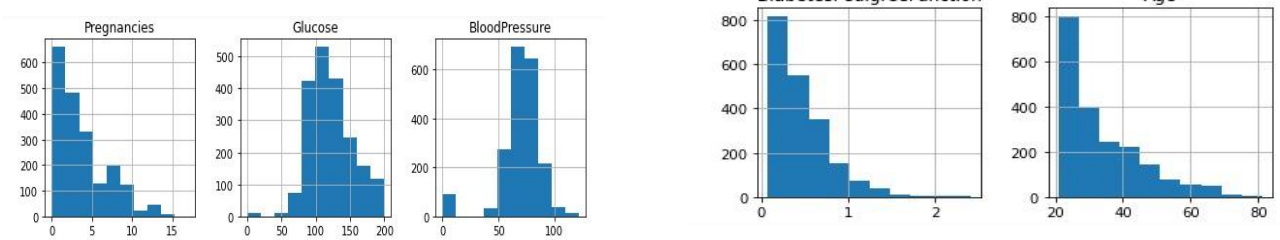

Fig 3: Number of people vs attribute measure

Fig 3 shows plots of various people with different attribute ranges which give a glimpse of each attribute's value.

3.4. Techniques

1. Logistic Regression: It is a simple yet effective algorithm aimed at binary classification problems like diabetes prediction. Here, it can be used to predict the chances of a person having diabetes based on the given parameters. We found that it works well when there are linear relationships between the features and the target variable. Some coefficients are obtained from the models which can measure and show the importance of each parameter on having influence on the target variable 'diabetic'. As per Fig. 7, the range of features in the trained dataset is normalized by feature scaling. Further, sigmoid function is applied on this dataset to classify the person as diabetic or not.

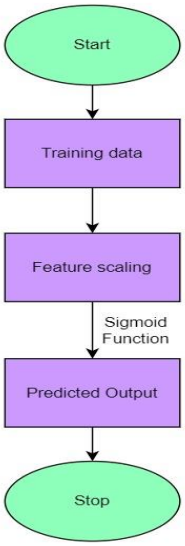

Fig 4: Logistic Regression

2. Support Vector Machines (SVM): It is a versatile algorithm that handles both linear and non-linear relationships in a dataset hence it can be beneficial for a dataset that has complex parameters. By using an appropriate kernel function (such as radial basis function), SVM can collect and transform complex patterns in the input features which can then be used for interpreting data easily. For diabetes prediction, SVM can be trained to find the optimal hyperplane that best separates individuals with diabetes from those without, in the multi-dimensional feature space defined by parameters like Glucose, BMI, and blood pressure. According to Fig. 5 acquired data is pre-processed and mapped on a hyperplane for classification of data points. Post mapping, the person is classified as diabetic or non-diabetic and performance analysis of the model was done.

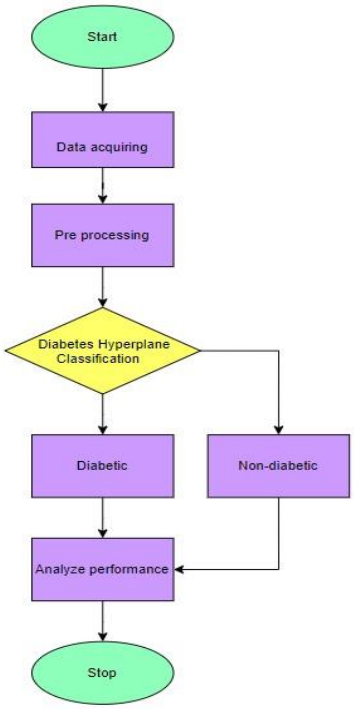

Fig 5: Support Vector Machine

3. Decision Trees: They are more based on intuitive models that make decisions based on a set of rules that can be derived from the input features. Every tree has a node that represents a feature, and there are branches which are possible feature values. For diabetes prediction, a Decision Tree is constructed to make splits based on parameters like Glucose level, Age, and BMI. The tree structure provides insights into which parameters are most influential in predicting diabetes.

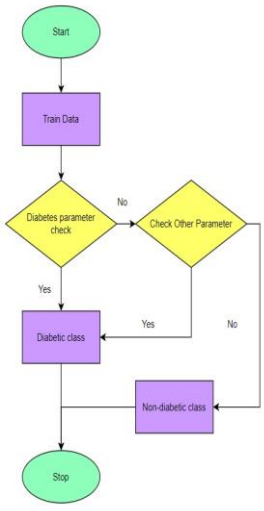

Fig 6: Decision Tree

4. Random Forest: It is an ensemble learning method that is a combination of multiple Decision Trees and is effective in improving accuracy and reducing overfitting. A random forest model is used when specified parameters have to be used as features and hence it can be constructive for diabetes(many parameters). Diverse patterns in the data can

be recorded using individual trees in the ensemble which is a collection of decision trees. Random Forest's ability to handle non-linear relationships and interactions among features makes it well-suited for this task. As per Fig. 7 multiple decision trees are built by considering all parameters and output is predicted on the basis of the majority of the result from every decision tree.

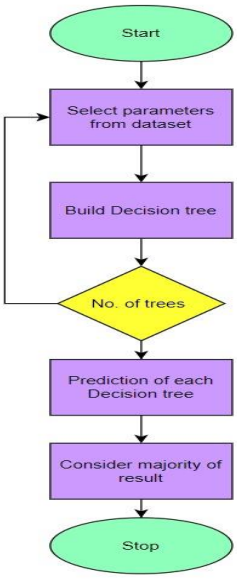

Fig 7: Random Forest

For most models dataset is divided into training and testing data and then various models which we have mentioned above are implemented on it where we train the models on the data and further this is evaluated on the testing data where we can assess the performance of the diabetes predicting model. Each model's parameters are fine-tuned using techniques like cross-validation to improve their performance. Consistent refining these models with change in data can then help with improving accuracy and reliability of the model in the real-world.

4. Results and Discussions

Before after applying various models on the data with different interpretations found during EDA we have presented the results in the form of confusion matrix and performance metrics.

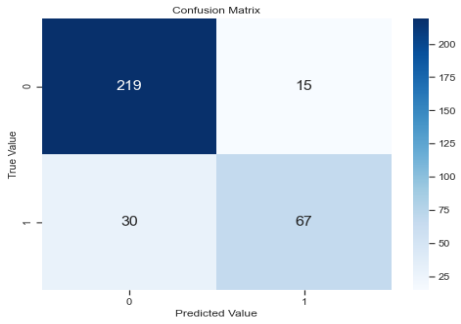

Fig 8: Confusion matrix for Clinical data using Gradient boost

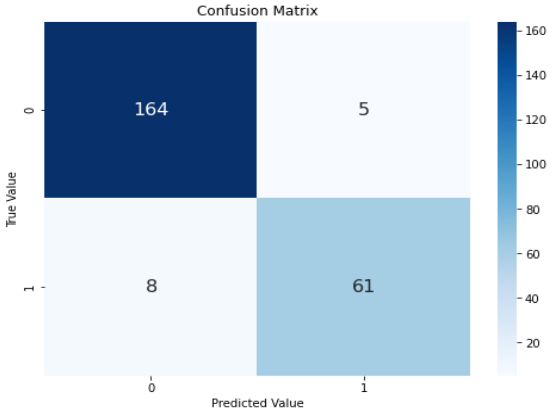

Fig 9: Confusion matrix for non-clinical data using gradient boost

From both Fig 8 and 9 we were able to see that the true positives offered by gradient boost are highest followed by true negatives and then false negatives and false positives. This is inline with the trend of most models but was the best of all confusion matrices. It proves the success of our model implementation.

Table 5: Performance metrics of Clinical factors

	LR	SVM	RFC	XGB	DT
Accuracy	0.822	0.810	0.979	0.903	0.976
Precision	0.722	0.775	0.981	0.898	0.972
Recall	0.598	0.539	0.931	0.775	0.922
F1 score	0.674	0.636	0.964	0.832	0.959
ROC Curve	0.760	0.735	0.966	0.868	0.961

Table 6: Performance metrics of Non-clinical factors

	LR	SVM	RFC	XGB	DT
Accuracy	0.878	0.895	0.966	0.945	0.971
Precision	0.898	0.887	0.982	0.953	0.971
Recall	0.935	0.976	0.970	0.970	0.988
F1 score	0.916	0.930	0.976	0.962	0.979
ROC curve	0.837	0.836	0.963	0.927	0.958

LR ~ Logistic Regression, SVM ~ Support Vector Machine, RFC ~ Random Forest Classifier, XGB ~ Extreme Gradient Boosting, DT ~ Decision Tree

It is clear that Random forest has the highest accuracy for both kinds of dataset and another revelation is that metrics for all models of non-clinical factors have had similar performance but clinical factors prove to be the best predictors when it comes to diabetes.

In this study, we explored the development of a robust diabetes predictor model based on several crucial features including Glucose, age, height, weight, BMI, blood pressure, in addition to gender. These were some clinical factors which we all know are important in making a good prediction but we have also explored a few techniques on non-clinical factors. Utilizing these parameters, we aimed to create an accurate and reliable prediction system for diabetes, a prevalent and critical health condition affecting millions worldwide.

The selection of these features was grounded in medical literature and expert opinions, recognizing their significant roles in diabetes risk assessment. Glucose levels are a fundamental marker, with elevated levels indicating potential diabetes risk. Additionally, age, weight, PhysicallyActive, BMI are widely acknowledged risk factors, and gender provides a more nuanced analysis, considering the biological difference between males and females.

5. Conclusion

In conclusion, we have found that Diabetes has had an impact on the health of various people and especially in India where we have a variety of food and a rigorous working schedule, it is difficult to find time to exercise corresponding to the unhealthy food that we have. So, this research paper aimed to develop a diabetes predictor model using four different machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forest. The objective was to determine the most effective method for predicting diabetes based on a set of input features. Through extensive experimentation and evaluation, some valuable insights have been gathered regarding the performance of these algorithms in the context of diabetes prediction.

After the deep study, analysis and implementation of the mentioned model, it was found that Random Forest emerged as the most superior method for predicting diabetes in this study. Random Forest showed exceptional accuracy, robustness, and generalizability when compared to Logistic Regression, SVM, and Decision Trees. The model has the ability to handle complex relationships within the data and reduce the chances of overfitting all the time while maintaining true accuracy makes it the method of choice for diabetes prediction in our research; especially the truthfulness of prediction matters in the healthcare industry.

In summary, the research helped in finding the best out of various methods in machine learning and it comes out to be Random Forest. This has led to an increase in accuracy. Moreover, with exclusive focus on feature selection and data selection we found that they play an important role in deciding the solidity of the predictor system. It can increase the inclusiveness of data driven models for patient care which will indirectly lead to advancement of predictive healthcare analytics that also affects. As the field of machine learning continues to evolve, the insights gained from this study pave the way for further research and innovation, ultimately leading to improved healthcare outcomes for individuals at risk of diabetes.

References

[1] Amelec Viloria, Yaneth Herazo-Beltran, Danelys Cabrera, Omar Bonerge Pineda, Diabetes Diagnostic Prediction Using Vector Support Machines, Procedia Computer Science, Volume 170, 2020, Pages 376-381, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.065.

[2] Rani, KM. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 294-305. 10.32628/CSEIT206463.

[3] Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019,Pages 292-299, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.

[4] Nongyao Nai-arun, Rungruttikarn Moungmai, Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science, Volume 69,2015, Pages 132-142, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.10.014.

[5] Kakoly, Israt Jahan, Md. Rakibul Hoque, and Najmul Hasan. 2023. "Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique" *Sustainability* 15, no. 6: 4930. https://doi.org/10.3390/su15064930

[6] Osama R. Shahin, Hamoud H. Alshammari, Ahmad A. Alzahrani, Hassan Alkhiri, Ahmed I. Taloba, A robust deep neural network framework for the detection of diabetes, Alexandria Engineering Journal, Volume 74, 2023, Pages 715-724,1110-0168, https://doi.org/10.1016/j.aej.2023.05.072.

[7] Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. 2019 1st International Informatics and Software Engineering Conference(UBMYK). doi:10.1109/ubmyk48245.2019.8965556

[8] E. Ismail, R. Ramdan, S. Mohamed, R. Yousri and M. S. Darweesh, "A Comparative Study of Diabetes Classification Based on Machine Learning," 2023 Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, 2023, pp. 598-603,doi:

10.1109/IMSA58542.2023.10217373.

[9] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin and M. K. Islam, "A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 2021, pp. 654-659, doi: 10.1109/ISPCC53510.2021.9609364.

[10] Varun Jaiswal, Anjli Negi, Tarun Pal, A review on current advances in machine learning based diabetes prediction, Primary Care Diabetes, Volume 15, Issue 3, 2021, Pages 435-443,ISSN1751-9918, https://doi.org/10.1016/j.pcd.2021.02.005.

[11] Oza A., Bokhare A. (2022). Diabetes Prediction Using Logistic Regression and K-Nearest Neighbor. In: Saraswat, M., Sharma, H., Balachandran, K., Kim, J.H., Bansal, J.C. (eds) Congress on Intelligent Systems. Lecture Notes on Data Engineering and Communications Technologies, vol 111. Springer, Singapore. https://doi.org/10.1007/978-981-16-9113-3_30

[12] F. M. Okikiola, O. S. Adewale and O. O. Obe, "An Ontology-Based Diabetes Prediction Algorithm Using Naïve Bayes Classifier and Decision Tree," 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG), Omu-Aran, Nigeria, 2023, pp. 1-11, doi: 10.1109/SEB-SDG57117.2023.10124491.

[13] Ratna Patil, Sharvari Tamane, Shitalkumar Adhar Rawandale, and Kanishk Patil. "A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus." *Int. J. Electr. Comput. Eng* 12, no. 1 (2022): 524-533.

[14] Wu, Han & Yang, Shengqi & Huang, Zhangqin & He, Jian & Wang, Xiaoyi. (2017). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 10. 10.1016/j.imu.2017.12.006.

[15] Zhou, H., Myrzashova, R. & Zheng, R. Diabetes prediction model based on an enhanced deep neural network. *J Wireless Com Network* 2020, 148 (2020). https://doi.org/10.1186/s13638-020-01765