ISSN : 2455-9679
SJIF Impact Factor : 6.008

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"DIABETES DISEASE PREDICTION USING RANDOM FOREST BASED MACHINE LEARNING MODEL"

Prachi Sharma [1], Dr. Akriti Jain [2]
[1]M. Tech Scholar, Department of CSE, LNCT, Bhopal (India)
[2]Associate Professor, Department of CSE, LNCT, Bhopal (India)

ABSTRACT

Medical decision support can improve healthcare processes by providing objective information about individual patients and it can be used to shift the analytic work from humans to computers. There are various disease occurred due to bad life style or any other reason, diabetes is one of them. Disease diagnosis is the most important health function. It can save lives if the disease is diagnosed before the usual or planned period. Machine learning-based classification methods can support the healthcare industry by enabling rapid and reliable disease diagnosis. Machine learning proved to be useful for detecting correlations in huge, complicated datasets. In this research work, used three machine learning techniques to predict diabetes disease Support Vector Machine (SVM), Random Forest (RF), and logistic regression (LR) model.

Key Words: *Diabetes disease, Machine learning, Classification, World health organization, Random forest, Support vector machines.*

I. INTRODUCTION

Diabetes is a chronic disease that requires frequent blood glucose testing. The World Health Organization's most recent statistics indicate that there are more than 500 million diabetic patients worldwide, and around 1.6 million people die each year due to diabetes and related disorders. The number of people living with diabetes may reach 780 million by 2045, according to reports from the International Diabetes Federation (IDF). To prevent or delay long-term health issues associated with diabetes, it is necessary to keep the sugar levels at the desired level. The hormone insulin, which is created in the pancreas, regulates the body's blood glucose level.

When blood glucose levels rise, the insulin production increases in the pancreas to counteract the rise [1]. This helps maintain the blood glucose level within the normal range in a healthy person. In type 2 diabetic patients, the body cannot produce enough insulin to adequately counteract the rise in blood glucose levels. Monitoring the level of glucose in a blood sample is the clinically accepted way for detecting diabetes. There is an increasing need for a non-invasive technique of monitoring diabetes because traditional diabetes detection is an invasive process. According to medical studies, small amounts of glucose are found in saliva, tears, sweat and urine. These biological samples offer the potential for non-invasive glucose level prediction in the body. Recent research has shown that the analysis of breath is a reliable non-invasive method to check glucose levels in the body [2].

There are no proper medications for diabetes cure until it is detected in its early stages. If diabetes is detected in its early stages, it can be managed easily through a healthy lifestyle as, illustrated in below figure. In early times, disease was diagnosed by manually reporting, which was an error-prone and an unreliable method. Generally machine learning

ISSN : 2455-9679
SJIF Impact Factor : 6.008

techniques are divided into some categories like supervised, unsupervised, and reinforcement learning, where supervised learning algorithms are practiced with labeled examples where the desired output is known. Based on the input, the algorithm learns by comparing its actual output and identifies its errors by which it modifies the model. It can also be applied by using historical data (known data) and predict future events. The training from the known dataset helps to learn the algorithm leading to an inferred function for the prediction of output values, and thus they can recommend targets for new input once trained. The algorithms build a model with the inputs and the desired outputs using training data with a set of examples.

Figure 1: Diabetes management.

Diabetes is divided broadly into two major types, type 1 diabetes and type 2 diabetes. Type 1 diabetes is characterized as a severe inadequacy or absence of releasing insulin by the pancreas due to an unknown disorder in the immune system, it affects people at a younger age more often, and it also can affect children. The primary treatment method for type 1 diabetes is insulin therapy. Type 2 diabetes is characterized by insulin resistance in which the body has insulin but does not utilize it well to regulate the blood sugar levels. This type is considered the most common one among diabetes patients. Recently, the incidence of type 2 diabetes has become alarmingly high, which is attributed to several reasons related to risk factors associated with diabetes, such as obesity, poor eating habits, lack of physical activity, smoking, and alcohol consumption.

Machine learning (ML) performs tasks autonomously without being explicitly programmed. IBM has been connected with ML for a long time. Arthur Samuel, an employee of IBM then, invented the term “machine learning” while researching the game of checkers program. ML is a process where computers learn from the input data for carrying out a certain task [21]. Algorithms involving the steps necessary for assigning simple tasks to the computers can be easily programmed without the learning needed by the computer. However, for advanced tasks, creating an algorithm manually is a bit complicated and challenging for a programmer. Rather, if the machine is allowed to create its own task list for performing the steps needed, the results would be more effective.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

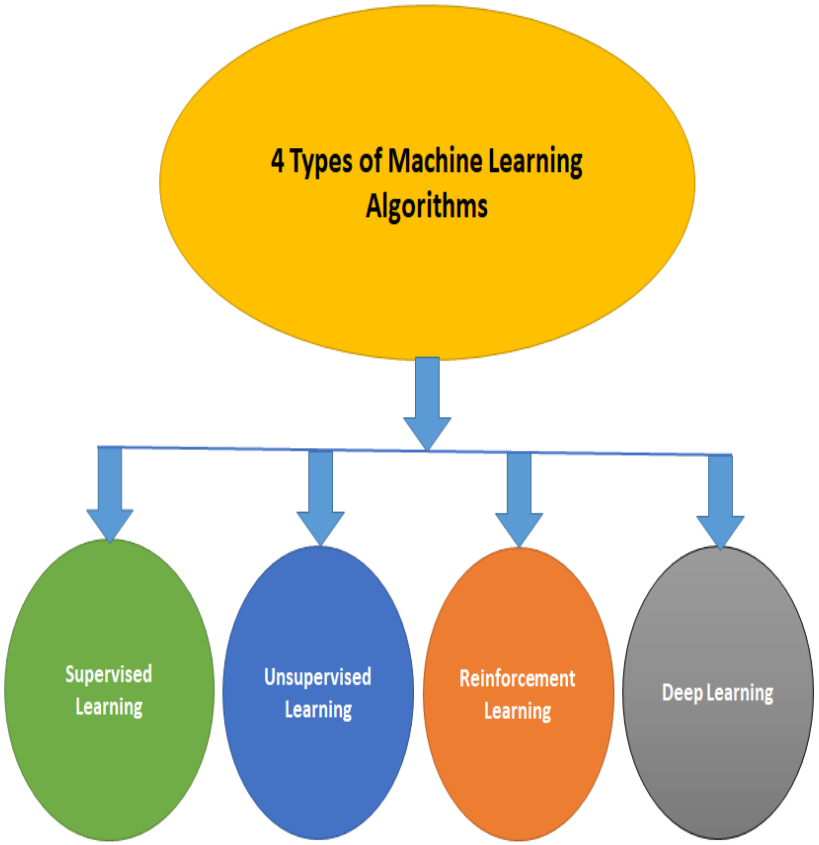

Figure 2: Types of machine learning algorithms.

The rest of this paper is organized as follows in the first section describe an introduction of about the diabetes disease overview, and machine learning. In section II we discuss about the literature work using machine learning techniques, In section III we discuss about the proposed model architecture of diabetes disease prediction system, In section IV we discuss about the experimental work, finally in section V we conclude the this research work and also suggest future directions.

II. Related Work

This study presents a proposed prediction model for diabetes that involves pre-processing techniques applied to the raw data, followed by the utilization of Ensemble Classifiers. The Ensemble Classifiers consist of a combination of catboost, LDA, LR, Random Forest, and GBC [1]. By employing pre-processing approaches and ensemble methodologies, they have achieved improved performance i.e. 90.62% accuracy. [2] In this research study, they developed machine learning (ML) and deep learning (DL) models to predict nocturnal glucose within the target range (3.9–10 mmol/L), above the target range, and below the target range in subjects with T1D managed with MDIs. The resulting models based on the DL and ML algorithms demonstrated high and similar accuracy in predicting target glucose (F1 metric: 96–98%) and above-target glucose (F1: 93–97%) within a 30 min prediction horizon. [3] According to the World Health Organization (WHO), some chronic diseases such as diabetes mellitus, stroke, cancer, cardiac vascular, kidney failure, and hypertension are essential for early prevention. One of the prevention that can be taken is to predict chronic diseases using machine learning based on personal medical record or general checkup result. The common prediction objective is to minimize the prediction error as low as possible. This research works covers machine learning methods discussion such as supervised learning, ensemble learning, deep learning, and reinforcement learning. [4] The proposed approach has achieved impressive performance. For the private dataset, the XGBoost algorithm with SMOTE achieved an accuracy of 97.4%, an F1 coefficient of 0.95, and an AUC of 0.87. For the

ISSN : 2455-9679
SJIF Impact Factor : 6.008

combined datasets, it achieved an accuracy of 83.1%, an F1 coefficient of 0.76, and an AUC of 0.85. To understand how the model predicts the final results, an explainable AI technique using SHAP methods is implemented. [5] In this research work, they address the critical issue of predicting survival in individuals with Diabetes-related complications, considering the interconnected nature of these complications. Non-communicable diseases (NCDs) such as Type-I and Type-II diabetes account for a significant global health burden, with its major complications such as cardiovascular disease as the leading cause. This research work explores a comprehensive analysis of survival prediction for Diabetes-related complications, utilizing six machine learning classification methods. Three methods, namely Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Neural Network (NN). [6] In this study, adaboost exhibited an AUC of 1 with overfitting, a binary classification model performance metric. Strategies to avoid overfitting include collecting more data, selecting relevant features, regularization techniques, cross-validation, early stopping, ensemble methods, and regular evaluation.

III. PROPOSED WORK

This paper has described the threats that computer viruses to research and development multi-user computer systems; it has attempted to tie those programs with other, usually simpler, programs that can have equally devastating effects. Many author were investigated how anti-virus software analyzes the infected file and shows pro-missing approach for malware detection in the future. To combat the never ending virus generation, the anti-virus software company should work closely with researchers to find potential approach that both work efficiency and accuracy.
Disease diagnosis is the most important health function. It can save lives if the disease is diagnosed before the usual or planned period. Machine learning-based classification methods can support the healthcare industry by enabling rapid and reliable disease diagnosis. As diabetes disease is difficult to diagnose, so it's a good time for doctors and patients. Here review the indicated machine learning and classification methods. The aim of the work is understanding whether a patient who has diabetes disease or not.

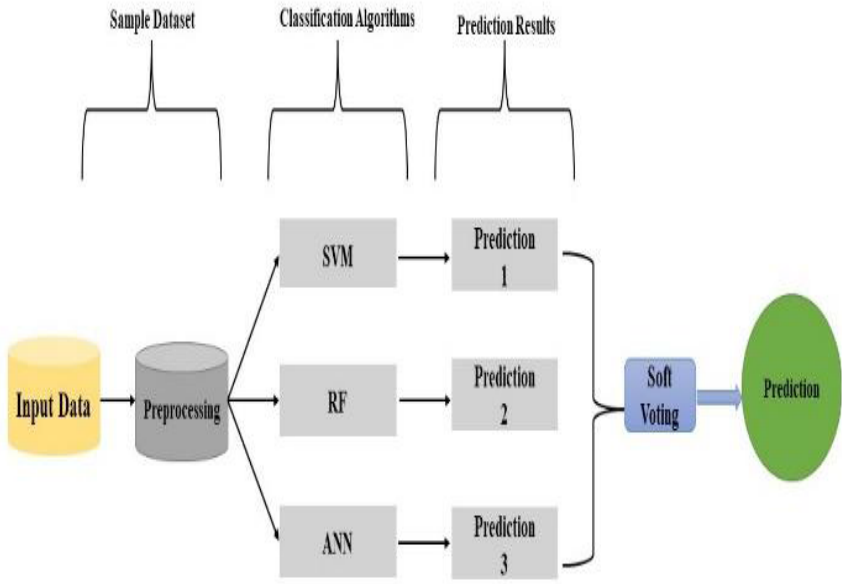

Figure 3: Proposed system flowchart for prediction of diabetes disease

These are the steps involved in the prediction of diabetes:
1. The system first receives the diabetes data set as input.
2. Based on the given symptoms, the diabetes predictor assists by predicting the presence of diabetes and generates the predicted results.
3. The diabetes monitor device assists in checking blood sugar levels and sends out
alerts based on them.
4. The user receives the awareness message to know about their health status.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

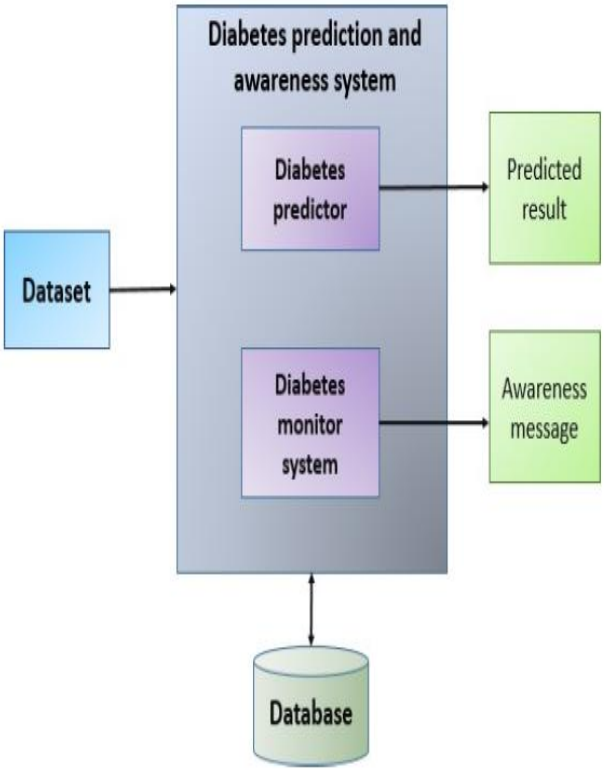

Figure 4: Proposed data architecture.

IV. EXPERIMENTAL WORK

In this chapter present the experimental result which is based on the machine learning techniques that predict the diabetes diseases is present or absent in a particular patient's. Predicting diabetes disease utilizing numerous machine learning algorithms like support vector machines, random forest, and logistic regression model, all these algorithms have been applied to the diabetes dataset available at kaggle datasets.

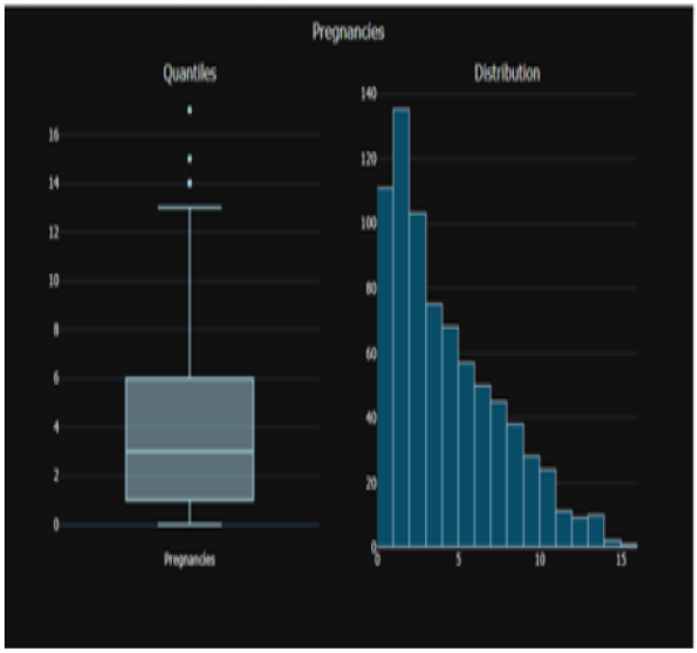

Figure 5: This picture shows the dataset data frame.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

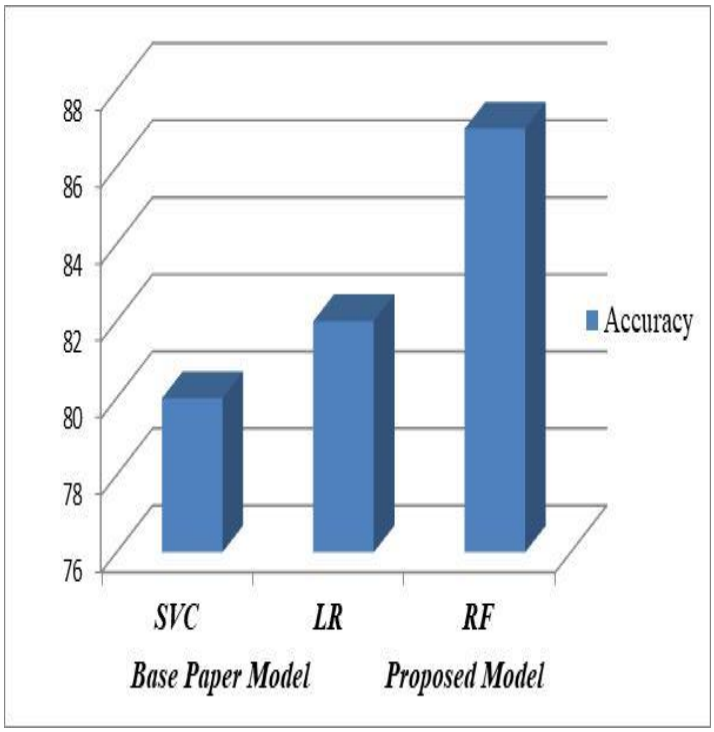

Figure 6: This picture represents the performance parameter value of accuracy between base paper based machine learning model and proposed machine learning based model.

V. CONCLUSION

Diabetes can significantly impact life expectancy and quality of life, making early prediction of this chronic disorder crucial for reducing long-term risks and complications. The performance metrics used for evaluation are as follows: Accuracy, Precision, Recall, and F1 Score. After applying all the ML models finally we found some performance parameters value and all models are compared with their respective performance parameters value, our result suggest that random forest algorithm gives better results than other techniques. The objective of this research work is to provide prediction using different SML algorithms. Using these algorithms and compared with each other to figure out which is the best. The algorithms Random Forest gives highest accuracy than other techniques or models. Future research can explore the integration of additional algorithms, such as deep neural networks, to further enhance accuracy and precision. Additionally, the use of swarm optimization techniques can be considered to optimize results. Application program development could also be incorporated to enhance the overall system.

REFERENCES

[1] Kok-Lim Alvin Yau, Yung-Wey Chong, “Reinforcement Learning Models and Algorithms for Diabetes Management”, IEEE Access, 2023, pp. 28391-28415.

[2] Navaneeth Bhaskar, Vinayak Bairagi, “Automated Detection of Diabetes From Exhaled Human Breath Using Deep Hybrid Architecture”, IEEE Access, 2023, pp. 51712-51723.

[3] Abdul Muiz Fayyaz, Muhammad Imran Sharif, “Analysis of Diabetic Retinopathy (DR) Based on the Deep Learning”, Information 2023, pp. 1-14.

[4] Huiqi Lu, Xiaorong Ding, “Digital Health and Machine Learning Technologies for Blood Glucose Monitoring and Management of Gestational Diabetes”, IEEE, 2022, pp. 1-19.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

[5] Yifei Su, Chengwei Huang, “Diabetes Mellitus Risk Prediction Using Age Adaptation Models”, Biomedical Signal Processing and Control, 2022, pp. 1-20.

[6] Chaitanya Krishna Suryadevara, “Diabetes Risk Assessment Using Machine Learning: A Comparative Study of Classification Algorithms”, International Engineering Journal For Research & Development, 2022, pp. 1-11.

[7] Y. Jeevan Nagendra Kumar, “Prediction of Diabetes using Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, 2019.

[8] A. Garrison, "Screening, diagnosis, and management of gestational diabetes mellitus," Am Fam Physician, vol. 91, no. 7, pp. 460-467, Apr 2015.

[9] W. W. Zhu, "High prevalence of gestational diabetes mellitus in Beijing: effect of maternal birth weight and other risk factors," Chin Med J, vol. 130, no. 9, pp. 1019-1025, May 5 2017.

[10] C. C. Martis R, "Treatments for women with gestational diabetes mellitus: an overview of Cochrane systematic reviews " Cochrane Db Syst Rev, REVIEW no. 8, 2018.

[11] J. E. Hirst, "Preventing childhood obesity starts during pregnancy," Lancet, vol. 386, no. 9998, pp. 1039-40, Sep 12 2015.

[12] J. E. Hirst, "GDm-health: development of a real-time smartphone solution for the management of women with gestational diabetes mellitus (GDM)," Bjog-Int J Obstet Gy, vol. 122, pp. 403-403, Apr 2015.

[13] L. Loerup, "A comparison of blood glucose metrics to assess the feasibility of a digital health system for management of women with gestational diabetes: the GDm-Health study," Diabetic Med, vol. 32, pp. 18-19, Mar 2015.

[14] P. A. Dyson, "GDm-Health Plus: Development of a remote behavioural lifestyle management system for women with gestational diabetes," Diabetic Med, vol. 35, pp. 171-171, Mar 2018.

[15] M. Peleg, "MobiGuide: a personalized and patient-centric decision-support system and its evaluation in the atrial fibrillation and gestational diabetes domains," User Model User-Adap, vol. 27, no. 2, pp. 159-213, Jun 2017.

[16] G. García-Sáez, "Patient-oriented computerized clinical guidelines for mobile decision support in gestational diabetes," Journal of Diabetes Science and Technology, vol. 8, no. 2, pp. 238-246, 2014.

[17] L. M. Garnweidner-Holme, "Designing and developing a mobile smartphone application for women with gestational diabetes mellitus followed-up at diabetes outpatient clinics in Norway," Healthcare-Basel, vol. 3, no. 2, pp. 310-323, Jun 2015.

[18] I. Borgen, "Smartphone application for women with gestational diabetes mellitus: a study protocol for a multicentre randomised controlled trial," Bmj Open, vol. 7, no. 3, Mar 2017.

[19] I. Borgen, "Effect of the pregnant plus smartphone application in women with gestational diabetes mellitus: a randomised controlled trial in Norway," Bmj Open, vol. 9, no. 11, Nov 2019.

[20] F. Debong, H. Mayer, and J. Kober, “Real-World Assessments of mySugr Mobile Health App”, Diabetes Technol Ther, vol. 21, no. S2, pp. S235-s240, Jun 2019.

[21] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning", IEEE, pp 942-928, 2018.

[22] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes", Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[23] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[24] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques", Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13.

[25] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach", IEEE Congress on Evolutionary Computation (CEC), 2018.