ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

Volume 11, Issue 2, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

IMPACT FACTOR: 7.583

| www.ijarasem.com | ijarasem@gmail.com | +91-9940572462 |

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 2, March 2024 |

Diabetes Prediction and Classification Using Machine Learning

Dr.N.NAVEENKUMAR, SHREE KANTH M, SUDHARSAN S, VIMALRAJ N

Associate Professor, Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

ABSTRACT: Diabetes is a serious and chronic condition. Diabetes can be detected early enough which can result in more effective treatment. This study also compares various classification models based on machine learning algorithms for predicting a patient's diabetic condition at the earliest feasible stage. After dataset balancing, classifiers' accuracy was compared. The prime objective of our research is to determine the early prediction of diabetes using the state of advanced MLA in one of the rural areas of North Kashmir. The data set employed for this experimentation was gathered from clinical professional. In the medical diagnosis, we used diabetes clinical data set with 403 instances and 11 attributes. The professionals (Prediabetes specialists) in the medical field have approved the features chosen for the early diagnosis of diabetes Prediction.

KEYWORDS: machine learning; diabetes; decision tree; random forest; support vector machine; k-nearest neighbor

I. INTRODUCTION

In many research studies, well-known machine learning techniques, including the Naïve Bayes classifier, support vector machines, decision trees, random forests, K-nearest neighbors, and logistic regression, have been widely used in diabetes classification [6]. The performance of these machine learning algorithms is mainly evaluated based on a benchmark PIMA Indian Diabetes dataset [6]. Most researchers provide a few steps of data preprocessing and hyperparameter tuning to increase the accuracy of their promising classifiers. For example, Zhao and Miao (2018) [7] conducted a comprehensive experiment to compare the accuracy of five popular machine learning techniques, namely, logistic regression, DNNs (deep neural networks), SVMs (support vector machines), decision trees, and the Naïve Bayes classifier, using the PIMA Indian dataset across several methods of data preprocessing, including imputation, scaling, and normalization, among others. In addition, the authors performed parameter optimization for each classifier and analyzed the features' effect to verify the relevance of features used in diabetes identification. This study revealed that scaling should be conducted for preprocessing. Although DNNs are the most accurate technique, they require a much longer run time and have more parameters to modify than SVMs and decision trees, which have a less reduced accuracy. Zou et al. (2018) [8] used five-fold cross-validation based on the PIMA Indian data and another dataset from a local hospital in Luzhou, China, to examine the accuracy of three classification methods (decision tree, random forest, and neural network). Principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) were also employed to reduce dimensionality. It was found that there was not much difference between the three algorithms. Nonetheless, the random forest was better than the others in some dimension-reduction methods as it uses all features, and mRMR was better than PCA. Nandhini A and Dharmarajan (2022) [9] focused on the accuracy of random forest (RF) algorithms in terms of various feature selection methods. Compared to other feature selection methods, the exhaustive feature selection with the random forest classifier and hyperparameter tuning using the grid search view gave the best result.

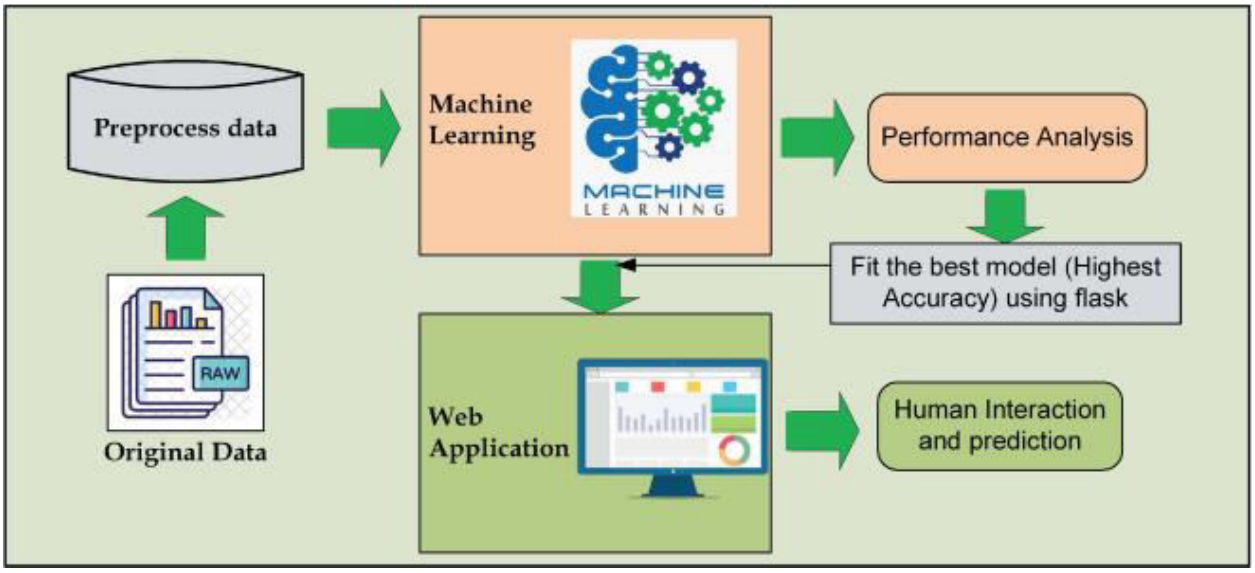

Fig 1: Machine learning based diabetes prediction

Several feature selection and construction methods can be used to identify interac-tions among important risk factors that improve the performance of diabetic classification models. Cheng et al. (2023) conducted regression tree analysis to identify the interactions among risk factors that contribute to glycated hemoglobin (HbA1c) values in type 2 diabe-tes mellitus. They found evidence suggesting that depression can be an important factor in certain subgroups of type 2 diabetes mellitus (T2DM). Using regression tree analysis, three pathways of multiple risk factors associated with poor glycemic control in T2DM patients were identified. Compared to other machine learning methods, the random forest algorithm was the best-performing method with a small set of features. In particular, the random forest algorithm achieved 84% accuracy, 95% area under the curve (AUC), 77% sensitivity, and 91% specificity.

II. LITERATURE REVIEW

Besides, we have also presented the IoT-based hypothetical diabetes self-monitoring system that uses BLE (Bluetooth Low Energy) devices and data processing in real-time. The latter technique used two applications: Apache Kafka (for streaming messages and data) and MongoDB (to store data). By utilizing BLE-based sensors, one can collect essential sign data about weight and blood glucose. These data will be handled by data processing techniques in a real-time environment. A BLE device will receive all the data produced by sensors and other necessary information about the patient that resides in the user application, installed on the cell phone. The raw data produced by sensors will be processed using the proposed approach to produce results, suggestions, and treatment from the patient's server-side.
The rest of the paper is organized as follows. In, the paper presents the motivations for the proposed system by reviewing state-of-the-art techniques and their shortcomings. It covers the literature review about classification, prediction, and IoT-based techniques for healthcare. highlights the role of physical activity in diabetes prevention and control. In, we proposed the design and architecture of the diabetes classification and prediction systems. discusses the results and performance of the proposed approach with state-of-the-art techniques. In, an IoT-based hypothetical system is presented for real-time monitoring of diabetes. Finally, the paper is concluded in, outlining the future research directions.

Health condition diagnosis is an essential and critical aspect for healthcare professionals. Classification of a diabetes type is one of the most complex phenomena for healthcare professionals and comprises several tests. However, analyzing multiple factors at the time of diagnosis can sometimes lead to inaccurate results. Therefore, interpretation and classification of diabetes are a very challenging task. Recent technological advances, especially machine learning techniques, are incredibly beneficial for the healthcare industry. Numerous techniques have been presented in the literature for diabetes classification.

Utilized data mining techniques, i.e., random forest, logistic regression, and naïve Bayes algorithm, to predict diabetes at the early or onset stage. They used 10-fold cross-validation and percentage split techniques for training purposes. They collected diabetic and nondiabetic data from 529 individuals directly from a hospital in Bangladesh through

questionnaires. The experimental results show that random forest outperforms as compared to other algorithms. However, the state-of-the-art comparison is missing and achieved accuracy is not reported explicitly.

III. METHODS

Generally, physical activity is the first prevention and control strategy suggested by healthcare professionals to diabetic or prediabetic patients. Among diet and medicine, exercise is a fundamental component in diabetes, cardiovascular disease, obesity, and lifestyle rescue programs. Nonetheless, dealing with all the fatal diseases has a significant economic burden. However, diabetes mellitus emerged as a devastating problem for the health sector and economy of a country of this century.

Recently, the international diabetes prevention and control federation predicts that diabetes can affect more than 366 million people worldwide. The disease control and prevention center in the US alarmed the government that diabetes can affect more than 29 million people. While these alarming numbers are continuously increasing, they will burden the economy around the globe. Therefore, researchers and healthcare professionals worldwide are researching and proposing guidelines to prevent and control this life-threatening disease. Sato presented a thorough survey on the importance of exercise prescription for diabetes patients in Japan. He suggested that prolonged sitting should be avoided and physical activity should be performed every 30 minutes.

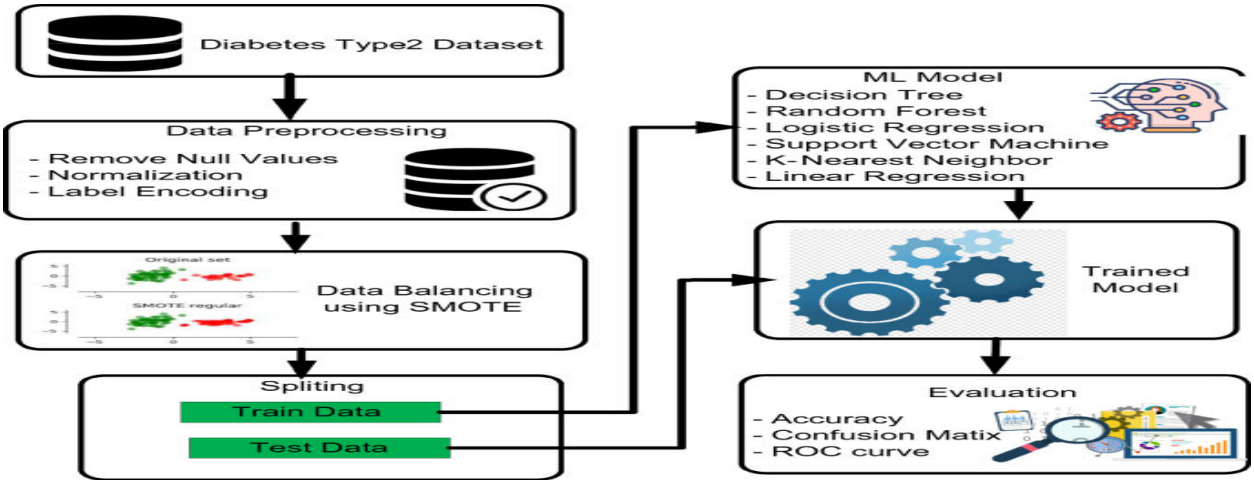

Fig 2: Machine Learning Based Diabetes Detection

Shows the multilayer perceptron classification model architecture where eight neurons are used in the input layer because we have eight different variables. The middle layer is the hidden layer where weights and input will be computed using a sigmoid unit. In the end, results will be computed at the output layer. Backpropagation is used for updating weights so that errors can be minimized for predicting class labels. For simplicity, only one hidden layer is shown in the architecture, which in reality is much denser.

IV. RESULT ANALYSIS

We collected the clinical dataset using the snow sampling technique by collaborating with a clinical diabetic professional. The collected dataset has 403 instances each with 11 attributes. The dataset does not contain any personal information such as the names of the person or their personal identification numbers in order to protect their privacy. However, the data is imbalanced. There are multiple techniques through which we can remove in balancing factors so that overall accuracy can be improved. The experimental study's dataset, which was constructed using clinical data in accordance with the endocrinologist's recommendations (diabetes specialists).

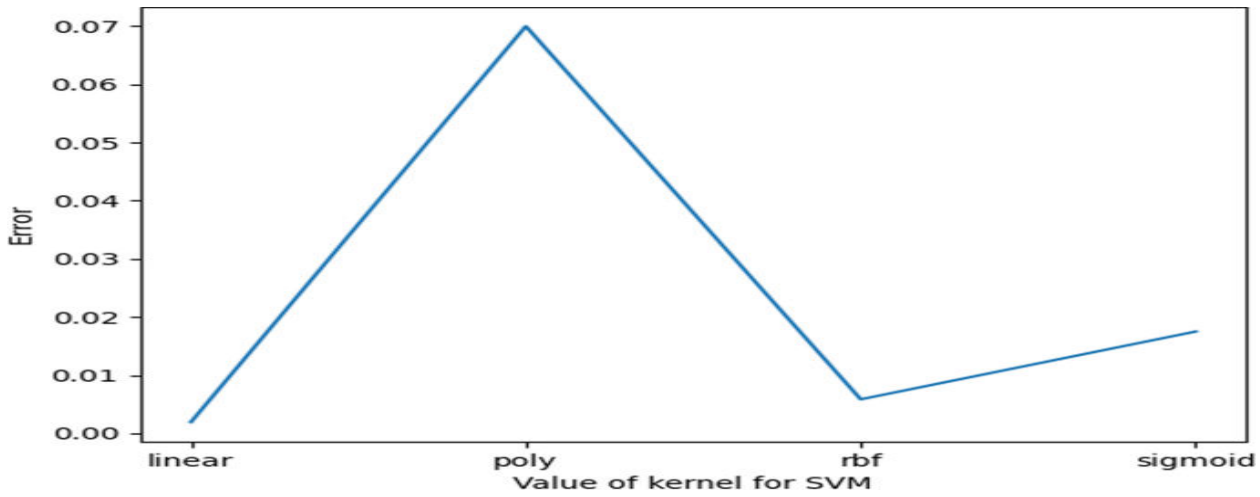

Fig 3: Result analysis of diabetes prediction

The clinical dataset was used to evaluate and test the suggested method. In the clinical dataset, there are many different types of disorders. The raw data are transformed into a data analysis file format for cleaning and extraction of features. The methodology presented in this article defines diabetic medicine. The patient is different from the optimum healthy patient.

V. CONCLUSION

In conclusion, this research presents new classification models that incorporate optimized hyperparameters and include the interaction of important risk factors affecting diabetes. The results reveal that, upon tuning the hyperparameters and including the interaction terms, the proposed models have better performance than those without interaction terms for all four techniques (decision tree, random forest, support vector machine, and K-nearest neighbor). Among the proposed models with interaction terms, random forest had the best performance classification, with 97.5% accuracy, 97.4% precision, 96.6% recall, and a 97% F1-score. The proposed models with interaction terms are more efficient than the models without interaction terms because we included interaction with important risk factors affecting diabetes, body mass index, and a family history of diabetes in the models. The findings from this research can be further developed into a program to effectively screen potential diabetes patients in the future.

Nevertheless, other attributes related to exercise, lifestyle (such as waist-to-height ratio), and dietary management (including protein, fat, and sugar intake control) have also been identified as important risk factors for diabetes

REFERENCES

1. Available online: **https://www.who.int/news-room/fact-sheets/detail/diabetes** (accessed on 29 April 2023).
2. Available online: **https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html** (accessed on 29 April 2023).
3. Griffin, P.; Rodgers, M.D. Type 1 Diabetes. National Institute of Diabetes and Digestive and Kidney Diseases. Available online: **https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-1-diabetes** (accessed on 14 April 2023).
4. Griffin, P.; Rodgers, M.D. Risk Factors for Type 2 Diabetes. National Institute of Diabetes and Digestive and Kidney Diseases. Available online: **https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes** (accessed on 14 April 2023).
5. Available online: **https://www.cdc.gov/diabetes/basics/risk-factors.html** (accessed on 29 April 2023).
6. Pacharawongsakda, E. *An Introduction to Data Mining Techniques*; Pearson Education: London, UK, 2014. [**Google Scholar**]
7. Wei, S.; Zhao, X.; Miao, C. A comprehensive exploration to the machine learning techniques for diabetes identification. In Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 5–8 February 2018; pp. 291–295. [**Google Scholar**] [**CrossRef**]
8. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus with Machine Learning Techniques. *Front Genet.* **2018**, *9*, 515. [**Google Scholar**] [**CrossRef**] [**PubMed**]

9. Sneha, N.; Tarun, G. Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data* **2019**, *6*, 13. [Google Scholar] [CrossRef]
10. International Statistical Classification of Diseases and Related Health Problems 10th Revision. Available online: https://icd.who.int/browse10/2019/en#/E10-E14 (accessed on 29 April 2023).

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com