International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

A Survey on Using Machine Learning to Predict Diabetes Early on

Satyendra Singh Rawat[1], Amit Kumar Mishra[2], Deepak Motwani[3]
[1]Department of Computer Science & Engineering, Swami Rama Himalayan University,
Swami Ram Nagar, Jollygrant, Dehradun, INDIA
[2,3]Department of Computer Science & Engineering, Amity University,
Bhind Road, Deen Dayal Nagar, Gwalior, INDIA

Abstract- Diabetes is a category of metabolic disease caused by a prolonged high blood sugar level. It is sometimes referred to as a chronic disease. If accurate early prediction is achievable, it can considerably lower the risk factor and severity of diabetes. Combining data mining methods with machine learning, a subsection of artificial intelligence, offers promise in the field of prediction. Data is widely available in the healthcare industry, and in order to improve prognosis, diagnosis, therapy, medication development, and healthcare in general, information must be extracted from it. Based on the World Health Organisation's 2014 report, diabetes is a type of chronic disease with the fastest global growth rates. To illustrate the widely used techniques for early diabetes detection—which are based on cutting-edge technologies including machine learning, cloud computing, etc.—we have reviewed a few significant pieces of literature in this study. The findings suggested that artificial intelligence-based methods are more effective in the early detection of diabetes in patients. Here, we used the Random Forest model to conduct an experiment using a diabetes dataset. First, the dataset is resampled and then used to train and test the Random Forest model. On all performance criteria, the Random Forest attained values above 96%.

Keywords: AI, Diabetes, Deep Learning, IoT, Machine Learning,

1. INTRODUCTION

It is commonly recognized that excess weight and abdominal fat accumulation pose significant risks for those with type 2 diabetes. Using physical examinations together to foresee the presence of type 2 diabetes is still debatable. The purpose of the research is to forecast, in adult Koreans, the fasting plasma glucose (FPG) level utilised in the determination of type 2 diabetes using a variety of measurements (Lee & Kim, 2016). Type 2 diabetes is closely linked to the hypertriglyceridemic waist (HW) phenotype (Lee & Kim, 2016). The diagnosis of diabetic retinopathy depends heavily on the detection of microaneurysms (MAs), which are thought to be the initial lesions in the condition (Zhou et al., 2017).

When making medical decisions, estimating the risk of long-term diabetic problems is crucial. The supervision of type 2 diabetes mellitus (T2DM) guidelines includes estimating the risk of cardiovascular disease (CVD) in order to start the right medication (Zarkogianni et al., 2018). New advances in wearable computing and artificial intelligence, combined with advancements in big data and wireless networking technologies like 5G networks, remedial big data analytics, and the Internet of Things (IoT), are making it possible to create and deploy creative diabetes monitoring apps and systems. It is essential to develop efficient techniques for the diagnosis and treatment of diabetics due to the chronic and systemic harm that these individuals experience (Chen et al., 2018). Hyperglycemia is a chronic disorder allied with diabetes mellitus. It can encourage a number of issues. Based on the increasing morbidity in recent times, 642 million people worldwide will have diabetes by 2040, which implies that one in ten adults will have the disease. Without a doubt, careful consideration is needed for this concerning number (Zou et al., 2018). Continuous glucose monitoring systems (CGMSs), which enable high sample rates for measuring a diabetic patient's blood glucose value, produce a significant amount of data. Machine learning approaches have the potential to utilise this data to infer future glycaemic concentration values, which could lead to better treatment optimization for diabetics and early prevention of severe hyperglycaemic or hypoglycaemic situations (Aliberti et al., 2019). Finding the underlying genes of an illness is crucial to understanding its genesis. One of the main issues facing human health is our understanding of the relationship between underlying genes and genetic disorders. Extensive and costly experimentation on a large number of putative candidate genes is necessary for the identification and connection of genes with the disease. Consequently, other quick and low-cost computational techniques that can determine the potential gene linked to an illness have been put forth (Sikandar et al., 2020). Diabetes, also known as chronic disease, is a collection of metabolic syndromes

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

carried on by insistently raised blood sugar levels. If a precise initial forecast can be made, it can considerably minimize the risk factors and severity of diabetes (Hasan et al., 2020).
Every year, hundreds of millions of people worldwide suffer from diabetes-related health issues. The clinical laboratory test results, physical features, and symptoms of patients are quantified in their available medical records. These data can be used for biostatistics analysis, which looks for patterns or traits that are currently missed by treatment (Le et al., 2021). One solution to the significant risk associated with diabetes is to diagnose the condition early. With so many characteristics to consider, predicting a patient's early start of diabetes is a challenging undertaking (Samreen, 2021). Diabetic retinopathy (DR) is a worsening condition that affects the eyes and is caused by high blood sugar levels in diabetes mellitus. For diabetic patients, particularly those in underdeveloped countries who are of working age, diabetic retinopathy is thought to be the primary cause of blindness (Atwany et al., 2022). An important prevalent post-translation modification (PTM) of proteins in creatures is N-linked glycosylation, which occurs when the protein's asparagine (N) amino acid is joined to the glycan. It is a component of almost every biological process and is linked to a number of illnesses in humans, including diabetes, cancer, influenza, coronavirus, and Alzheimer's. Thus, figuring out N-linked glycosylation places will help us comprehend the glycosylation system and process (Alkuhlani et al., 2022). One of the most potent and quickly expanding technologies in use today is health information technology. As pathological testing and medical visits can be costly and time-consuming, this technology is primarily employed for sickness prediction and fast medicine delivery (Himi et al., 2023).

2. REVIEW OF LITERATURE

- The goal of this study was to find out how type 2 diabetes is linked to the HW phenotype in adult Koreans and how well other phenotypes, which are made up of different anthropometric parameters and TG levels, can predict type 2 diabetes (Lee & Kim, 2016). Non-microaneurysm (MA) data vary widely, and gathering non-MA training sets is a complex task that takes time and may lead to class imbalance issues. In order to identify MA, an unsupervised classification technique linked to sparse PCA was created in this work (Zhou et al., 2017). The aim of the research is to examine the application of advanced ML methods to the creation of customised models that can forecast the likelihood of either fatal or non-fatal cardiovascular disease (CVD) occurrence in individuals with type 2 diabetes mellitus (T2DM). Using information from 560 T2DM patients' medical records, the models were tested; the best discrimination performance was 71.48% in terms of AUC (Zarkogianni et al., 2018). This presents the 5G-smart diabetes system, which generates complete sensing and investigation of diabetic patients by combining cutting-edge technologies like wearable 2.0, ML, and big data. The outcomes of the experiment demonstrate how well our system can offer patients individualised diagnosis and therapy recommendations (Chen et al., 2018). Applying machine learning to several elements of medical health has become possible due to its rapid progress. The prediction of diabetes mellitus was done in this study using neural networks, random forests, and decision trees. Based on all the attributes combined, the results indicated that random forest prediction could achieve the maximum correctness (ACC = 0.8084) (Zou et al., 2018).

- The majority of methods in the literature get their prediction model from previous patient samples, which restricts the system's applicability and necessitates lengthy calibrations. The aim of this exercise is to inspect prediction models that were trained on a vast and diverse set of patient glucose signals and then use them to envisage upcoming glucose levels in a newly diagnosed patient (Aliberti et al., 2019). This exercise attempts to identify and evaluate a few unique computational approaches for identifying disease-associated genes. The outcomes show that several computational approaches with sophisticated feature sets perform better than earlier state-of-the-art methods, attaining up to 93.8% precision, 93.1% recall, and 92.9% F-measure (Sikandar et al., 2020). Because there are few labelled data points and outliers or mislaid values in the diabetes datasets, it is very hard to make a reliable and accurate diabetes prediction. Our suggested ensemble classifier is the best-performing classifier, outperforming competitors by 2.0% (Hasan et al., 2020).

- The objective of the study is to generate an ML pipeline that can create the most illustrative feature subset in the smallest possible size and provide the greatest accuracy prediction for the start of diabetes (Samreen, 2021). To preserve high-accuracy findings while lowering the requirements for measurement, storage, and computing, feature selection (FS) is adopted. Our proposal involves the use of Grey Wolf Optimizer (GWO) and Adaptive Particle Grey Wolf Optimisation (APGWO) to choose features using a wrapper approach, taking into account the nondeterministic polynomial-time complex characteristic of FS. Predictive accuracy can be increased while using significantly fewer features, according to computational research (Le et al., 2021). The study proposes the categorisation and detection of retinal fundus images by reviewing and analysing the most recent advances in deep learning techniques in unsupervised,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

self-supervised, and Vision Transformer settings (Atwany et al., 2022).

- Nonetheless, a number of previous literary works make assumptions about diabetes without taking incomplete and unbalanced evidence into account. To fix these problems, we present in this study a useful probabilistic ensemble classification method for the diabetes dataset to handle missing values in the wrong class (PE_DIM). This method can quickly fix missing values and improve the accuracy of classification (Jia et al., 2022). The difficulties in identifying glycosylation sites experimentally make machine learning crucial for glycosylation site prediction. This study presents a unique N-linked glycosylation predictor called PUStackNGly, which is based on stacking ensemble machine learning and bagging positive-unlabeled (PU) learning. On a separate dataset, the suggested PUStackNgly performs better than the current N-linked glycosylation prediction methods (Alkuhlani et al., 2022). We have concentrated on categorising individuals with diabetes in this study by utilising a variety of sensors, including glucose, an accelerometer, an ECG, and breathing sensors. We examined the accuracy of single-sensor and multiple-sensor diabetes prediction. According to the performance curves, three sensor combinations—glucose, accelerometer, and ECG—converge more quickly than four sensor combinations while maintaining the same level of accuracy (Site et al., 2023).

- The paper offers a wearable device-based prediction system called "MedAi" that uses ML algorithms to foresee the risk of a number of diseases, including dyslipidemia, hypertension, ischemic heart disease, etc. With an accuracy of 99.4%, according to experiments conducted with our dataset (Himi et al., 2023). I introduced an ANN model that trains a classifier using the guided stochastic gradient descent algorithm. Our results show that the guided ANN works better than the regular one. It converges faster and is more accurate at classifying patients with diabetes (Ram et al., 2023) in the class-2 and class-3 datasets by at least 1% to 1.5. This work uses a convolutional neural network (CNN) model to suggest a unique method for detecting diabetic retinopathy. The experimental results show that, when compared to its competitor's models, the proposed CNN model achieves higher scores of 96.85%, 99.28%, 98.92%, 96.46%, and 98.65% for accuracy, sensitivity, specificity, precision, and f1 score (Ali et al., 2023).
- Approximately 537 million people worldwide suffer from diabetes, and by 2045, that figure is likely to rise to 783 million. One of the main problems associated with diabetes that might result in lower limb amputation is diabetic foot ulcers (DFU). Due to the fast progression of DFU, prompt action is necessary to avert the catastrophic outcomes of amputation and associated comorbidities. This article presents a novel method for accurately classifying photos of diabetic foot ulcers (DFUs) using deep neural networks and machine learning (Toofanee et al., 2023). Early and accurate DR detection could reduce the amount of harm lost. This work presents the development of a new IoTDL-DRD model for retinal fundus picture DR classification and detection. A comprehensive comparative analysis revealed that the suggested approach performed better (Palaniswamy & Vellingiri, 2023). Individuals living in remote locations are not receiving the appropriate medical care. Therefore, in order to lower the death rate and give rural residents access to medical care, this study project suggested an automated eHealth cloud system for early diabetes detection (Sharma et al., 2023).

3. EXPERIMENTAL RESULTS

Pima Indians Diabetes Database

The initial source of this dataset is the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset's objective is to analytically forecast the probability that an individual has diabetes based on particular diagnostic criteria that are included. Following some limitations, these examples were selected from a larger database. Most notably, all of the patients at this clinic are 21-year-old Pima Indian women.
Several medical predictor factors make up the datasets, whereas Outcome serves as the only target variable. The patient's age, BMI, insulin level, past pregnancy history, and others are predictor variables. The dataset contains two classes of data instances: normal patients (0) and diabetes patients (1). Normal class contains 500 data instances and diabetes class contains 268 data instances. The initial distribution of data instances is given in Fig. 1.

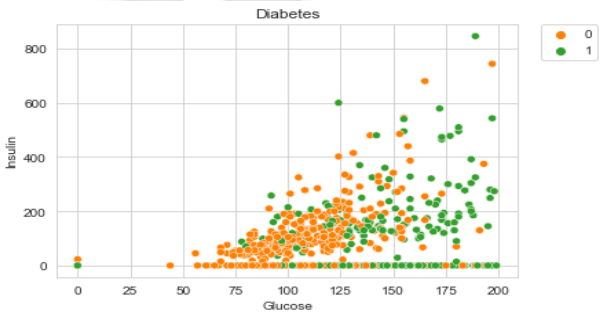

Figure 1. Initial data distribution into classes (0: Normal and 1: Diabetic)

Resampling the dataset

The dataset is resampled using the SMOTE-ENN (Muntasir Nishat et al., 2022) Synthetic Minority Oversampling Technique with Edited Nearest Neighbour method. The ENN

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

method is used to clean the dataset by removing noisy instances, outliers, etc. from the dataset. The SMOTE methods are used to construct diabetes data instances by linearly interpolating between a diabetes data instance and one of its k-k-nearest neighbours data instances. Figure 2 provides the resampled dataset's data instances. Here, green ovals represent instances of the diabetes class data, while yellow ovals represent instances of the normal patient's class data.

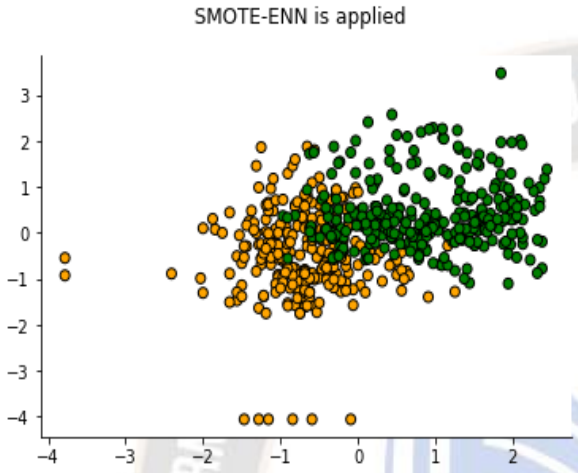

Figure 2. The distribution of data instances of the resampled dataset.

Training and Testing of Random Forest Model

As an ensemble model, the Random Forest (RF) (Breiman, 2001) is superior to the standard Machine Learning model. In this instance, the dataset is separated into a test set and a training set. After training on the training set, the RF is evaluated on the test set. Table 1 displays the experiment's outcomes.

Table 1. Experimental results of Forest model

	Precision	Recall	Accuracy	F1-score	G-mean	AUC
RF	97%	97%	96.5%	97%	96.4 %	96.4 %

4. CONCLUSION

Early disease identification lowers the major risk factor. Additionally, it lowers the costs associated with treating diabetes. Diabetes prediction systems nowadays are more effective since they rely on Internet-based technologies like IoT and cloud computing, but they are still limited to urban regions due to lower internet speed. Hence, systems based on ML and deep learning are becoming extensively used and well-liked. The machine learning model we have chosen in this case, Random Forest, was trained and tested using data related to diabetes. Various indicators were used for the performance examination, and RF received 96% on all of them.

REFERENCES

[1] Ali, G., Dastgir, A., Iqbal, M. W., Anwar, M., & Faheem, M. (2023). A Hybrid Convolutional Neural Network Model for Automatic Diabetic Retinopathy Classification From Fundus Images. IEEE Journal of Translational Engineering in Health and Medicine, 11, 341–350. https://doi.org/10.1109/JTEHM.2023.3282104

[2] Aliberti, A., Pupillo, I., Terna, S., MacIi, E., Di Cataldo, S., Patti, E., & Acquaviva, A. (2019). A Multi-Patient Data-Driven Approach to Blood Glucose Prediction. IEEE Access, 7, 69311–69325. https://doi.org/10.1109/ACCESS.2019.2919184

[3] Alkuhlani, A., Gad, W., Roushdy, M., & Salem, A. B. M. (2022). PUStackNGly: Positive-Unlabeled and Stacking Learning for N-Linked Glycosylation Site Prediction. IEEE Access, 10, 12702–12713. https://doi.org/10.1109/ACCESS.2022.3146395

[4] Atwany, M. Z., Sahyoun, A. H., & Yaqub, M. (2022). Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey. IEEE Access, 10, 28642–28655. https://doi.org/10.1109/ACCESS.2022.3157632

[5] Chen, M., Yang, J., Zhou, J., Hao, Y., Zhang, J., & Youn, C. H. (2018). 5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds. IEEE Communications Magazine, 56(4), 16–23. https://doi.org/10.1109/MCOM.2018.1700788

[6] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516–76531. https://doi.org/10.1109/ACCESS.2020.2989857

[7] Himi, S. T., Monalisa, N. T., Whaiduzzaman, M. D., Barros, A., & Uddin, M. S. (2023). MedAi: A Smartwatch-Based Application Framework for the Prediction of Common Diseases Using Machine Learning. IEEE Access, 11, 12342–12359. https://doi.org/10.1109/ACCESS.2023.3236002

[8] Jia, L., Wang, Z., Lv, S., & Xu, Z. (2022). PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance Missing Values. IEEE Access, 10, 107459–107476. https://doi.org/10.1109/ACCESS.2022.3212067

[9] Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2021). A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a

Metaheuristic. IEEE Access, 9, 7869–7884. https://doi.org/10.1109/ACCESS.2020.3047942

[10] Lee, B. J., & Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on Machine Learning. IEEE Journal of Biomedical and Health Informatics, 20(1), 39–46. https://doi.org/10.1109/JBHI.2015.2396520

[11] Palaniswamy, T., & Vellingiri, M. (2023). Internet of Things and Deep Learning Enabled Diabetic Retinopathy Diagnosis Using Retinal Fundus Images. IEEE Access, 11, 27590–27601. https://doi.org/10.1109/ACCESS.2023.3257988

[12] Ram, A. A., Ali, Z., Krishna, V., Nishika, N., & Sharma, A. (2023). A Guided Neural Network Approach to Predict Early Readmission of Diabetic Patients. IEEE Access, 11, 47527–47538. https://doi.org/10.1109/ACCESS.2023.3275086

[13] Samreen, S. (2021). Memory-efficient, accurate and early diagnosis of diabetes through a machine learning pipeline employing crow search-based feature engineering and a stacking ensemble. IEEE Access, 9, 134335–134354. https://doi.org/10.1109/ACCESS.2021.3116383

[14] Sharma, S. K., Zamani, A. T., Abdelsalam, A., Muduli, D., Alabrah, A. A., Parveen, N., & Alanazi, S. M. (2023). A Diabetes Monitoring System and Health-Medical Service Composition Model in Cloud Environment. IEEE Access, 11, 32804–32819. https://doi.org/10.1109/ACCESS.2023.3258549

[15] Sikandar, M., Sohail, R., Saeed, Y., Zeb, A., Zareei, M., Khan, M. A., Khan, A., Aldosary, A., & Mohamed, E. M. (2020). Analysis for Disease Gene Association Using Machine Learning. IEEE Access, 8, 160616–160626. https://doi.org/10.1109/ACCESS.2020.3020592

[16] Site, A., Nurmi, J., & Lohan, E. S. (2023). Machine-learning-based diabetes prediction using multi-sensor data. IEEE Sensors Journal, 1–1. https://doi.org/10.1109/JSEN.2023.3319360

[17] Toofanee, M. S. A., Dowlut, S., Hamroun, M., Tamine, K., Petit, V., Duong, A. K., & Sauveron, D. (2023). DFU-SIAM a Novel Diabetic Foot Ulcer Classification With Deep Learning. IEEE Access, 11, 98315–98332. https://doi.org/10.1109/access.2023.3312531

[18] Zarkogianni, K., Athanasiou, M., & Thanopoulou, A. C. (2018). Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication. IEEE Journal of Biomedical and Health Informatics, 22(5), 1637–1647. https://doi.org/10.1109/JBHI.2017.2765639

[19] Zhou, W., Wu, C., Chen, D., Yi, Y., & Du, W. (2017). Automatic Microaneurysm Detection Using the Sparse Principal Component Analysis-Based Unsupervised Classification Method. IEEE Access, 5, 2563–2572. https://doi.org/10.1109/ACCESS.2017.2671918

[20] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00515