IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 2, February 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 2, February 2023 ||

| DOI:10.15680/IJIRSET.2023.1202099 |

Suvey Paper on Prediction of Diabetes Disease using Machine Learning Technique

Nahid Noor Hussan[1], Prof. Suresh. S. Gawande[2]

M. Tech. Scholar, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal, India[1]

Guide, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal, India [2]

ABSTRACT: The global diabetes epidemic causes complications such as neuropathy, nephropathy, and retinopathy. Over time, diabetics may experience nerve damage throughout their bodies. Pain, tingling or numbness, and loss of feeling in the hands, and legs are common symptoms in these people. It's also more prone to skin illnesses including bacterial and fungal infections. If remain untreated, cuts and blisters can develop into dangerous infections that take a long time to cure, and severe damage may demand toes or leg amputation. In existing method, the classification and prediction accuracy is not so high. In this paper, we have studied a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

KEYWORDS*:* Diabetic dataset, Classification, Machine Learning

I. INTRODUCTION

Disease prognosis is one of the most important tasks when designing medical diagnostic software. In a variety of applications, including clinical diagnosis, machine learning methods are successfully used. Machine learning algorithms can greatly help solve health-related challenges by providing a predictive analytics system, which can help clinicians diagnose and diagnose diseases at an early stage. Diabetic neuropathy is a diabetes complication that leads to nervous system damage [1, 2]. Neuropathy tends to happen when nerves in the body are damaged by dominant levels of cholesterol or glucose in the blood. It can affect any nerve throughout the body, with symptoms being widespread. The common signs include burning sensations, muscle cramping or pain, hypersensitivity to touch, and significant foot problems including ulceration, inflammation, infections and joint complaints. The disease's irreversible progression inevitably leads to limb amputations. Diabetic nephropathy is a kind of progressive disease of the kidney that can occur in individuals with diabetes [3, 4]. The main signs are loss of appetite, itchy and dry skin, and swelling of the arms and legs. These patients have a higher risk of renal failure than the general population. Over time, the only treatment choices are dialysis or a kidney transplant. Type 1 diabetes is triggered by insufficient insulin synthesis and insulin does not function for persons experiencing type 2 diabetics who are affected by these neuropathy and nephropathy complications. Diabetes patients with type 1 are often diagnosed in childhood. But people with type 2 diabetes are usually diagnosed over 30- year-old [5]. Therefore, identifying the symptoms, causes and progressive complications is easier in type 2 than in type 1 categories. To improve disease diagnosis, the data science and machine learning platform combine multiple data sources into a unified framework.

II. LITERATURE REVIEW

Jyoti Agarwal et al. [1], in addition to blindness, kidney failure, and heart attacks, diabetes can also result in death. In 2019, there were 463 million diabetics, according to the International Diabetes Federation. This number will rise by 578 million by 2030, reaching 700 million by 2045, if predictions are correct. The ten nations with the highest rates of diabetes in 2019 include Indonesia, according to a 2020 article published by the Ministry of Health of the Republic of Indonesia. To identify the type of diabetes, experts must be able to do so. Many people who are examined have a disease that can be described as severe due to their delay in discovering their disease. To avoid severe conditions, technology for diabetes detection is necessary. Doctors can use it to quickly and accurately diagnose diseases in today's medical environment. As a result, we can use machine learning to prevent death by creating an artificial intelligent model that can predict diabetes disease. To determine which algorithm is best suited for diabetes prediction, we compare the KNN and

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 2, February 2023 ||

| DOI:10.15680/IJIRSET.2023.1202099 |

Naive Bayes algorithms. The study concluded with a comparison of two k-Nearest Neighbor algorithms and the Naive Bayes algorithm for using supervised machine learning to predict diabetes based on a number of health attributes in the dataset. The Naive Bayes algorithm performs better than the KNN algorithm in our experiments and when evaluated with the Confusion Matrix.

Anuj Mangal et al. [2], a group of metabolic disorders leads to diabetes mellitus, a common condition in which blood sugar levels remain extremely high for an extended period of time. It harms a lot of the body's systems because it affects various organs, particularly the nerves and blood vessels. It is possible to control such a disease and save lives through early prediction. Using machine learning methods, this study primarily investigates a variety of disease-related risk factors to achieve its objective. By building predicting models from diagnostic medical datasets gathered from diabetic patients, machine learning techniques provide efficient results for extracting knowledge. Predicting diabetic patients can benefit from learning from such data. On adult population data, we use four well-known machine learning algorithms—Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), and C4.5 Decision Tree (DT)—to predict diabetic mellitus in this study. According to our experiments, the C4.5 decision tree outperformed other machine learning methods in terms of accuracy.

Nicole D'Souza et al. [3], diabetes is a metabolic condition that affects many people worldwide. Every year, its incidence rates are alarmingly rising. Numerous vital organ complications caused by diabetes may result in death if left untreated. For timely treatment, which can stop the disease from progressing to such complications, early diabetes detection is critical. For the non-invasive detection of diabetes, RR-interval signals, also known as heart rate variability (HRV) signals, which are derived from electrocardiogram (ECG) signals, can be utilized effectively. A method for using deep learning architectures to classify normal HRV signals from those of diabetes is presented in this study. For the purpose of extracting complex temporal dynamic features from the input HRV data, we make use of convolutional neural networks (CNN), long short-term memory (LSTM), and their various combinations. For classification, these features are sent to a support vector machine (SVM). In comparison to our prior work that did not make use of SVM, our CNN-LSTM architecture performance improved by 0.03% and 0.06%, respectively. With an extremely high 95.7% accuracy, the proposed classification system can assist clinicians in diagnosing diabetes using ECG signals.

Yogita Dubey et al. [4], diabetes Mellitus is one of the most serious diseases, and many people have it. Age, obesity, inactivity, hereditary diabetes, bad diet, high blood pressure, and other factors can result in type 2 diabetes. Diabetes increases a person's risk of heart disease, kidney disease, stroke, eye problems, nerve damage, and other conditions. In hospitals, it is common practice to use a variety of tests to gather the necessary data for a diabetes diagnosis, after which the patient receives the appropriate treatment. Healthcare sectors benefit greatly from Big Data Analytics. Databases in the healthcare sector are massive. With the help of big data analytics, large datasets can be studied to uncover hidden data and patterns that can be used to extract knowledge from the data and predict outcomes accordingly. The classification and prediction accuracy of the current method is not particularly high. For better diabetes classification, we have proposed a diabetes prediction model in this paper that takes regular factors like glucose, BMI, age, insulin, and other variables into account. When compared to the existing dataset, the new dataset improves classification accuracy. In addition, a pipeline model for diabetes prediction was imposed with the intention of increasing classification accuracy.

S.M. Tahsin Zaman et al. [5], diabetes is one of the most common and persistent diseases in the world today. Because there isn't enough insulin produced, the human body has more glucose in the blood. Therefore, in order to combat and prevent the disease's potential risks, early diabetes detection is now essential. Using various machine learning models (Decision Tree, K-Nearest Neighbors, Naive Bayes, Random Forest) to analyze diabetic patients and detect diabetes in human bodies, we present a method for classifying diabetics in this paper. The Diabetes Dataset, or Pima Indian Diabetes (PIMA) Dataset, is used to evaluate the proposed methodology. We begin by preparing the data for the experiment using the various preprocessing methods. Each model runs the extracted training data through it. We use the sensitivity, precision, f1-score, and overall accuracy measurements to evaluate the classifiers' performance. However, the Random Forest model achieved the highest score in the experiment, with an accuracy of 86%. With the help of the "Internet of Things," we hope that these classification techniques can be combined with real-time data to create real-time devices for health care applications.

Binhe Chen et al. [6], have Logistic Regression and Gradient Boosting Machine (GBM) to build predictive models using the most recent records of 13,309 Canadian patients aged 18 to 90 and their laboratory data (age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein). The ability of these models to discriminate was evaluated using the area under the receiver operating characteristic curve

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 2, February 2023 ||

| DOI:10.15680/IJIRSET.2023.1202099 |

(AROC). To increase sensitivity, or the proportion of Diabetes Mellitus patients that the model correctly predicted, we employed the adjusted threshold method and the class weight method. We also compared these models to Decision Tree and Random Forest, two other learning machine methods.

III. MACHINE LEARNING TECHNIQUE

Machine learning assists individuals and organizations in reaching their rapidly increasing medical requirements, improving services, and lowering costs. It also aids medical professionals in more accurately diagnosing and treating disease. As the world gets clever by the day, MLA techniques that make things easier for patients can lead to more effective, holistic care strategies that control the speed of their disease attack and lead to better health [7].

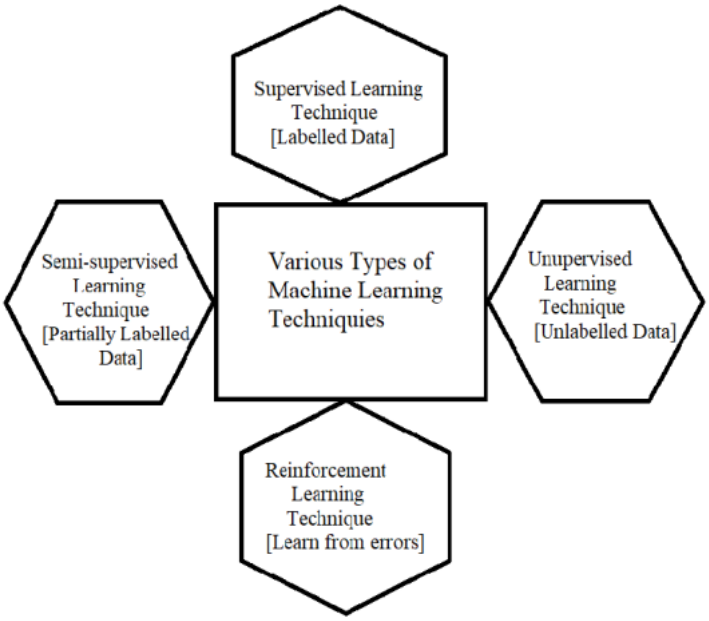

Figure 1: Various Types of ML Techniques

Such technology has the following different types: Supervised, unsupervised, semi-supervised and reinforcement learning. In the upcoming sections, the approaches are briefly described and depicted in the following Figure 1.

Supervised Learning:- The process of feeding labelled data to a machine learning model is known as supervised learning. The clearly labelled dataset is usually empirical data gained from experience. The optimum classification or regression algorithms are chosen depending on the application. Using supervised learning to predict diabetes and its complication have been demonstrated to be beneficial in many studies in this sector, and this work does so as well. The benefits of supervised learning include the efficiency with which the outcomes can be assessed. As a result, there are more reports on supervised learning in the field. In DPN prediction, a set of data with well-known classes is segmented generating data to train and validation. The categories data for training are submitted to a supervised MLA. The training models are evaluated using the validation set, which comprises the testing data. The success rate of test data provides an objective indication of a strategy's ability to forecast disease. In this study, statistical methods were used to find the components of the DPN and DN. However, the focus of this research is on supervised learning [8, 9].

Unsupervised Learning:- If getting labels is difficult or there isn't enough understanding of the data, unsupervised learning is employed. Furthermore, the lack of labels makes it tough to set targets for the classifier. As a result, assessing if the results are correct is challenging. Regardless, there are a number of tactics that can be employed to obtain outcomes that lead to a greater comprehension of the data. Clustering techniques like k-means clustering, KNN, Hierarchical clustering, and other clustering approaches find groupings of related items based on some of their characteristics. To put it another way, this method assists in the finding of data patterns. Cut Fiarni et al. developed a diabetic complication condition prediction model in Indonesia and discovered the critical features linked with it. For unsupervised learning, this study employed without labelled data to create a grouping that correlates with several of the consequence disorders. The

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 2, February 2023 ||

| DOI:10.15680/IJIRSET.2023.1202099 |

authors arrive at the best number of groups using a classic methodology for discovering the ideal volume of clusters from input, the k-means approach with the Elbow strategy. Examine the performance of the classifying and cluster analysis. In terms of identifying features and subfeatures, the classification technique surpassed clustering when compared to clustering [10]. Researchers compared GraphBased MLAs (GBMLS), a new clustering strategy, to conventional ML strategies for DPN. The findings show that the novel procedure outperformed other methods and produced the greatest results [4].

Semi-supervised Learning:- Unsupervised and supervised learning are combined in this type of learning. It uses a combination of labelled and unclassified/unlabelled data. This article provides a detailed overview of clinical data studies that use deep learning technologies. The authors give a comprehensive review of clinical data investigations involving deep learning technologies. To begin, a variety of medical data (such as clinical photographs, patient records, laboratory findings, vitals, and demographic census) is examined, and specifics of public health census data are provided based on clinical data aspects. With a better grasp of deep learning architecture and a better understanding of the predictions, based on these data. It also contributes to the conclusion that collaboration with other methodologies could lead to higher clinical analysis accomplishments [11]. This work investigated a datamining evolved forecasting methodologies using a way of gauging using the Minimum Description Length (MDL) principle for discretization. By combining data mining methodologies with reasonable accuracy, the semi-supervised learning strategy for diabetes prediction identifies crucial factors using clusters and classifiers [12].

Reinforcement Learning:- In order to construct the T2DM detection model, the Q-learning method from the Reinforcement Learning (RL) was used in the PIMA database. The authors focused on RL, specifically Q-learning, which is a trial-and-error technique for which agent trains in a dynamic situation using feedback through its own actions and experiences. The agent receives a mapping mechanism from a collection of facts to the ultimate consequences after experiencing a sufficient number of events. The trained model is used to complete this learning process. After the agent has been trained, the resulting projection approach is applicable to new test data to find patterns and predict outputs. The suggested Qlearning model is compared to existing different classifiers such as KN-Neighbour and DT in predicting whether or not an individual has T2D. According to the research, the KNN classifier scored beneath average, estimated with an accuracy of 74.49 % and a precision of 64.29 %. The DT classifier, from the other side, achieved an accuracy of 75.51%, recall was 54.55%. The proposed Q-learning model outperforms these ML classifiers, including an accuracy of 84%, precision of 100 %, and recall of 59.3 percent [13].
The goal of this study was to assist t2dm patients in increasing their physical activity levels. 27 T2DM patients were given a smartphone-based pedometer and a personalised physical activity regimen by the researchers. Patients received short message service messages once per day to once a week to encourage physical activity. A Reinforcement Learning technique was used to customize messages in order to increase each participant's compliance with the activity routine. Here, scientists built a smartphone app that gathers the amount of physical activity performed by patients and operates in the background of their smartphones. These records were sent to a central server. When compared to control policies, patients in the classifier group reported a larger reduction in blood glucose levels (glycated haemoglobin [HbA1c]), and prolonged involvement resulted in greater decreases in blood glucose levels. The study comes to a conclusion. Mobile phone apps mixed with an RL approach can improve exercise adherence in diabetics.
This strategy can be utilised to improve glucose tolerance and health in big populations of diabetics [14]. This is created and tested as a RL-based Evolutionary Fuzzy Rule-Based System (RLEFRBS) for diabetes screening. Creating a Rule Base (RB) and optimising rules are part of the suggested model. The first RB is formed using quantitative data and no preliminary rules; after understanding the rules, the confidence measure is used to remove unneeded rules. Then, in the preceding sections, unneeded requirements are removed to make the rules clearer and easier to understand. Lastly, the RB is constructed using a Genetic Algorithm (GA) to pick an acceptable subset of the rules. RLEFRBS' performance is increased by using Reinforcement Learning to fine-tune membership functions and change weights (RL). To deal with unwrapped occurrences, it also uses an effective rule extendible technique. The BioSat and Pima Diabetes repositories were used to evaluate RLEFRBS (BDD). The trials show that the suggested model produces an RB that is more compact, interpretable, and accurate, making it a promising alternative for diabetes diagnosis. This work proposes a one-of-a-kind FRBS to attain maximum accuracy and interpretability. The final RB has a high level of accuracy, according to the proposed model's performance evaluation findings. Furthermore, the collected data suggest that the RB closely matches real-world patient data, implying that it is capable of modelling diabetes behavior and reflecting knowledge in patient records [15].

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 2, February 2023 ||

| DOI:10.15680/IJIRSET.2023.1202099 |

IV. CONCLUSION

Diabetes is metabolic disease that arises due to high blood glucose level in body. Insulin is not sufficient enough in body of diabetic patients to regulate the sugar level. Further several other diseases also arise from diabetes which is hazardous to health. It is necessary to detect such a serious health issue as early as possible Diabetes is cause of various diseases in human body. To make diabetes diagnosis easier for Physicians, there have been several methods employed. The diabetic data set is tested with selected classification algorithm.

REFERENCES

[1] Jyoti Agarwal, Anu Sharma, Namit Gupta, P.R. Lakshmi Eswari and Satyanadha Sarma Samavedam, "Diabetes Predication Analysis using Supervised Machine Learning Algorithm", 11th International Conference on System Modeling & Advancement in Research Trends (SMART), IEEE 2022.

[2] Anuj Mangal and Vinod Jain, "Performance analysis of machine learning models for prediction of diabetes", 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), IEEE 2022.

[3] Nicole D'Souza, Kunjal Shah, Pranav Singh, "Diabetes Detection Using Machine Learning Algorithms", IEEE Bombay Section Signature Conference (IBSSC), IEEE 2022.

[4] Yogita Dubey, Pushkar Wankhede, Tanvi Borkar, Amey Borkar and Kajal Mitra, "Diabetes Prediction and Classification using Machine Learning Algorithms", IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), IEEE 2021.

[5] S.M. Tahsin Zaman, Subrata Kumer Paul, Rakhi Rani Paul and Md. Ekramul Hamid, "Detecting Diabetes in Human Body using Different Machine Learning Techniques", International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), IEEE 2021.

[6] Binhe Chen, Maosong Yan, Hongchuan Zhong and Bingwei He, "Prediction Model of Diabetes Based on Machine Learning", 5th Asian Conference on Artificial Intelligence Technology (ACAIT), IEEE 2021.

[7] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, "Prediction of Diabetes using Machine Learning Classification Algorithms", International Journal of Scientific & Technology Research, Volume 9, Issue 01, January 2020.

[8] Muhammad Azeem Sarwar, 2Nasir Kamal, 3Wajeeha Hamid,4Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2019.

[9] Rao G.A., Syamala K., Kishore P.V.V., Sastry A.S.C.S. ., "Deep convolutional neural networks for sign language recognition", 2018, International Journal of Engineering and Technology(UAE) ,Vol: 7, Issue 5, pp: 62 to 70.

[10] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang' "Predicting Diabetes Mellitus With Machine Learning Techniques", Springer, 2018.

[11] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neuro computing, vol. 237, pp. 350–361, May 2017.

[12] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[13] Reddy S.S., Suman M., Prakash K.N. ., "Micro aneurysms detection using artificial neural networks", 2018, Lecture Notes in Electrical Engineering ,Vol: 434 ,Issue 3, pp: 409 to 417.

[14] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15, 104–116, 2017.

[15] Majid Ghonji Feshki and Omid Sojoodi Shijan, "Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network", International paper on IEEE 2016.

[16] L. Hermawanti, "Combining of Backward Elimination and Naive Bayes Algorithm To Diagnose Breast Cancer", Momentum, vol. 11, no. 1, pp. 42-45, 2015.

[17] O.S. Soliman, E. Elhamd, "Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine", IEEE 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details