IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction by Classification Using Machine Learning and Deep Learning Techniques

[1]Siya Srivastava, [2]Shubha Mishra, [3]Shruti Srivastava

[1,3]Student, [2]Assistant Professor
[1,2,3]Computer Science and Engineering Department,
[1,2,3]Babu Banarasi Das Institute of Technology & Management, Lucknow, India

Abstract : Diabetes is a serious metabolic illness that can negatively impact every system in the body. The risk factors for this condition are Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. Diabetic nephropathy, heart stroke, and other illnesses are all made more likely by undiagnosed diabetes. Millions of people around the world are afflicted by this illness. To live a healthy life, it is crucial to catch diabetes early. Due to the significant increase in diabetes cases, this disease is a cause for concern on a global scale. The standard procedure in hospitals is to obtain the data needed for a diabetic diagnosis by a variety of tests, and based on that diagnosis, the proper therapy is given. This large volume of data collected from the patient has some hidden useful information, which can be used to prognosis, diagnose and treat the disease.

The revolutionary advancement in the field of Artificial Intelligence (AI) has opened the door to finding useful insights from the large dataset collected from various sources. All the sectors (healthcare, finance, Industrial sector, etc.) have heavily benefitted from it. Machine learning, which creates algorithms capable of learning patterns and decision-making rules from data, is one area where artificial intelligence has a greater impact. In order to extract knowledge from data, machine learning algorithms have been integrated into data mining pipelines, where they can be used in conjunction with conventional statistical techniques. The healthcare sector greatly benefits from Machine learning. Databases in the healthcare sector are very vast. By analysing large datasets using ML, one can learn from the data and make accurate predictions about the future by uncovering hidden patterns and information. The categorization and prediction accuracy of the current approach is not very good.

In this study, we suggested a diabetes prediction model that combines a few extrinsic factors that cause diabetes in addition to more common parameters like glucose, body mass index (BMI), age, insulin, etc. Compared to the old dataset, the new dataset improves classification accuracy. Additionally, a pipeline model for diabetes prediction was imposed with the goal of enhancing classification accuracy.

IndexTerms **- Diabetes, Machine learning, Analytics, artificial intelligence.**

INTRODUCTION

In the modern world, diabetes is a common term and a major problem in both industrialised and developing nations.[1] Glucose can enter the bloodstream from food thanks to the pancreas' production of the hormone insulin in the body. Diabetes is a condition in which there is insufficient production of that hormone due to pancreatic dysfunction. Diabetes can cause coma, renal and retinal failure, pathological destruction of pancreatic beta cells, diabetes dysfunction, cerebral vascular dysfunction, peripheral vascular diseases, sexual dysfunction, joint failure, weight loss, ulcers, and pathogenic effects on immunity.[2] According to research on diabetes patients, the prevalence of diabetes among adults (aged over 18) increased from 4.7% to 8.5% between 1980 and 2014, respectively, and is rising quickly in second and third-world nations[3]. Another statistical study [4] demonstrates the severity of diabetes. They found that there are 500 million diabetics globally, and that figure will rise to 25% and 51% in 2030 and 2045, respectively.

Although diabetes cannot be permanently cured, it can be managed and prevented if a reliable early prognosis is achievable. Since the distribution of classes for all attributes is not linearly separable, predicting diabetes is a difficult problem. Early diabetes diagnosis can result in more effective therapy. The standard procedure in hospitals is to obtain the data needed for a diabetic diagnosis by a variety of tests, and based on that diagnosis, the proper therapy is given. This large volume of data collected from the patient has some hidden useful information, which can be used to prognosis, diagnose and treat the disease. Over the past 40 years, the field of artificial intelligence (AI) has made significant advances in computer science and many of its application areas.[5], [6] Although the field of artificial intelligence hasn't entirely fulfilled to the expectations set in the 1970s and 1980s, its contributions to knowledge representation, modelling, automated reasoning, planning, and learning are remarkable. The AI branch of machine learning, which creates algorithms capable of learning patterns and decision rules from data, has recently received a lot of

attention.[7] Neural networks, deep learning, classification and association rules, support vector machines, and text mining pipelines are some of the algorithms that are entirely unique to the field. Other algorithms, like decision trees, naive bayes, logistic regression, and random forests, are borrowed from other fields. These techniques are frequently included into analytics pipelines that enable knowledge extraction from data in the form of clear models and useful recommendations for supporting decision-making. Such pipeline engineering is frequently referred to as data mining. Additionally, by building on already existing risk prediction calculators and combining them with the data available at a specific clinical site, data mining tactics can be used to create new predictive models that will help with illness management and treatment.

The goal of this project is to create a model that can predict a patient's chance of developing diabetes with the highest degree of accuracy. As a result, we will employ various machine learning and deep learning classification methods, to identify diabetes at an early stage.

LITERATURE REVIEW

The importance of deep learning techniques can be realised by their application in various fields, the health sector is one of them. In order to predict various diseases with high accuracy, many researchers are now focusing on deep learning algorithms. Additionally, the performance of treatment has changed noticeably as a result of the use of deep learning in the medical industry. A few representative works have been discussed as follows.

J. Wang, et al. demonstrated that deep learning may be utilised successfully to create an automated system for BAC identification in mammograms in order to identify and evaluate patients with diabetes risks. To distinguish between BAC and non-BAC, they created a 12-layer convolutional neural network and used a pixel-wise, patch-based method for BAC identification. They employed both calcium mass quantification analysis and free-response receiver operating characteristic (FROC) analysis. According to the FROC analysis, a degree of detection comparable to that of human specialists is achieved by the deep learning approach[8].

S. N. Pasha, et al. have shown that deep learning algorithms have better accuracy than machine learning algorithms. The accuracy of the ANN model was 85.24% for predicting CVD, which is almost 3 to 4% higher than SVM, KNN and Decision tree[9].

Quan Zou, et al. have reported Predicting Diabetes Mellitus With Machine Learning Techniques. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross-validation was used to examine the models. In order to verify the universal applicability of the methods, we chose some methods that have better performance to conduct independent test experiments. We randomly selected 68994 healthy people and diabetic patients' data, respectively as the training set. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. In this study, we used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used[10].

M. K. Hasan, et. al. reported diabetes prediction using the ensembling of different machine learning classifiers. They proposed a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed. From all the extensive experiments, they proposed ensembling classifier is the best-performing classifier with the sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC as 0.789, 0.934, 0.092, 66.234, and 0.950 respectively which outperforms the state-of-the-art results by 2.00 % in AUC[11].

A. Dagliati, et. al. reported the use of machine learning techniques to predict the complications related to diabetes. This work showed how data mining and computational methods can be effectively adopted in clinical medicine to derive models that use patient-specific information to predict an outcome of interest. Predictive data mining methods may be applied to the construction of decision models for procedures such as prognosis, diagnosis and treatment planning, which—once evaluated and verified—may be embedded within clinical information systems. Developing predictive models for the onset of chronic microvascular complications in patients suffering from T2DM could contribute to evaluating the relation between exposure to individual factors and the risk of the onset of a specific complication, to stratifying the patients' population in a medical centre with respect to this risk, and to developing tools for the support of clinically informed decisions in treatment[12].

MATERIAL AND METHODS

Quantitative research was performed in this study. The aim of the study was to accurately predict the presence of diabetes. For this purpose, we have employed a quantitative method. We have utilized the secondary dataset obtained from the PIMA Indian dataset, which was also used by other researchers.[13], [14], [15] Two classification models (two and five classes) were developed. First, a binary classification system was built to classify the patients based on with or without diabetes. Then, a multi-classification (five-class) system was built to measure the disease's severity.

The outlines of creating the model for the prediction of diabetes diseases -

3.1 Data Pre-processing

It is quite evident that pre-processing is an important aspect of machine learning.[16] First of all, we have done data cleaning. Then we checked whether the dataset has any missing values. Then we checked for the outliers in the dataset using a boxplot, plot and counterplot. Later removed, the outliers using Inter quartile range (IQR). further, we checked the unique values of the columns (features) and checked their correlation with the target variable. All the features were included in the model as they all have a strong correlation with the target variable. Now, this data was used for generating binary classification.

3.2 Data Description

We have utilized the secondary dataset obtained from the PIMA Indian dataset. The output contains two labels with no diabetes (0) and with diabetes (1). The data for training and testing was split into 80:20 and then output parameters were calculated.

3.3 Data transformation

The data must be in the right format for the data mining process. Data transformation or standardization is therefore required. Data redundancy is decreased. It is the process of altering data's configuration, values, and data structure. When one or more attributes are rescaled to have a mean value of 0 and a standard deviation of 1, the process is known as data normalisation. Data transformation is therefore necessary in order to have the same impact on all attributes. Data quality is raised as well as data organisation. This work uses min-max normalisation to transform Data for several methods.

RESULTS AND DISCUSSION

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database.

After using all these patient records, we are able to build a machine learning model (random forest – best one) to accurately predict whether or not the patients in the dataset have diabetes or not along with that we were able to draw some insights from the data via data analysis and visualization.

The proposed project compares various machine learning and deep learning models and it turns out that Random Forest Classifier gave out around 88% of accuracy.

CONCLUSION

In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

ACKNOWLEDGEMENT

I want to thank everyone who was involved for their knowledge, aid, and support in all areas of our study as well as for their assistance in authoring the manuscript.

REFERENCES

[1] A. Misra et al., "Diabetes in developing countries," J. Diabetes, vol. 11, no. 7, pp. 522–539, Jul. 2019, doi: 10.1111/1753-0407.12913.

[2] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Oct. 2017, pp. 1–5. doi: 10.1109/ICCNI.2017.8123815.

[3] The Emerging Risk Factors Collaboration, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies," The Lancet, vol. 375, no. 9733, pp. 2215–2222, Jun. 2010, doi: 10.1016/S0140-6736(10)60484-9.

[4] P. Saeedi et al., "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," Diabetes Res. Clin. Pract., vol. 157, p. 107843, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[5] J. H. Holmes, Ed., Artificial intelligence in medicine: 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015 ; proceedings. Cham: Springer, 2015.

[6] N. J. Ford, "Artificial intelligence: The very idea," Int. J. Inf. Manag., vol. 7, no. 1, pp. 59–60, Mar. 1987, doi: 10.1016/0268-4012(87)90008-9.

[7] D. Han et al., "Trends in biomedical informatics: automated topic analysis of JAMIA articles," J. Am. Med. Inform. Assoc., vol. 22, no. 6, pp. 1153–1163, Nov. 2015, doi: 10.1093/jamia/ocv157.

[8] J. Wang et al., "Detecting Cardiovascular Disease from Mammograms With Deep Learning," IEEE Trans. Med. Imaging, vol. 36, no. 5, pp. 1172–1181, May 2017, doi: 10.1109/TMI.2017.2655486.

[9] F. Fatima, A. Jaiswal, and N. Sachdeva, "Heart Disease Prediction Using Supervised Classifiers," SSRN Electron. J., 2022, doi: 10.2139/ssrn.4121817.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[11] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[12] A. Dagliati et al., "Machine Learning Methods to Predict Diabetes Complications," J. Diabetes Sci. Technol., vol. 12, no. 2, pp. 295–302, Mar. 2018, doi: 10.1177/1932296817706375.

[13] R. Krishnamoorthi et al., "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," J. Healthc. Eng., vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/1684017.

[14] U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," IEEE Access, vol. 10, pp. 8529–8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

[15] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning based diabetes prediction," Prim. Care Diabetes, vol. 15, no. 3, pp. 435–443, Jun. 2021, doi: 10.1016/j.pcd.2021.02.005.

[16] M. A. Hogo, "A proposed gender-based approach for diagnosis of the coronary artery disease," SN Appl. Sci., vol. 2, no. 6, p. 1060, Jun. 2020, doi: 10.1007/s42452-020-2858-1.