Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Effective Approach for Early Detection of Diabetes by Logistic Regression through Risk Prediction

K. Thangarajan

Associate Professor, Department of Electrical and Electronics Engineering, RVS College of Engineering and Technology, Coimbatore, India

E-mail: thangarajansk@gmail.com

Abstract

Heart disease, cancer, renal failure, eye damage, and blindness are just some of the complications that may result from uncontrolled diabetes. Scientists are inspired to develop a Machine Learning (ML) approach for diabetes forecasting. To improve illness diagnosis, medical personnel must make use of ML algorithms. Different ML algorithms for identifying diabetes risk at an early stage are examined and contrasted in this research. The goal in analysing diabetes prediction models is to develop criteria for selecting high-quality studies and synthesising the results from several studies. Nonlinearity, normality, correlation structure, and complexity characterise the vast majority of medical data, making analysis of diabetic data a formidable task. Algorithms based on machine learning are not permitted to be used in healthcare or medical imaging. Early diabetes mellitus prediction necessitates a strategy distinct from those often used. Diabetic patients and healthy individuals may be separated using a risk stratification approach based on machine learning. This study is highly recommended since it reviews a variety of papers that may be used by researchers working on diabetes prediction models.

Keywords: Early prediction, diabetic disease, machine learning, gradient boost classifier, risk factor

1. Introduction

The inability of the body to properly digest blood glucose (also known as blood sugar) is one of the primary symptoms of diabetes, a metabolic condition. Hyperglycemia is a hallmark of the condition, and it occurs because of a lack of insulin or an inability for insulin to do its job. Diabetes mellitus, more often known as diabetes, is a worldwide epidemic.

DOI: https://doi.org/10.36548/jaicn.2022.3.008
Received: 20.08.2022, received in revised form: 17.09.2022, accepted: 06.10.2022, published: 15.10.2022

Specifically, diabetes is a condition in which either insulin synthesis or insulin action is compromised. There are about 220 million persons infected with this disease [1-3]. High blood sugar from diabetes mellitus may lead to a number of health problems if it is not managed, including metabolic abnormalities and weight loss, eye diseases, renal disease, brain damage, and cardiovascular disease [4]. The prevalence of type 2 diabetes, which is more severe, has increased in recent years. Research in the area of diagnosis may benefit from the data and information offered by healthcare clinics. Type 2 diabetes mellitus is among the most frequently diagnosed illnesses worldwide. Pregnancy frequency, age, sex, blood pressure, heart rate, and other factors may all be used to estimate a woman's risk of developing cardiovascular disease. Figure 1 shows review of logistic regression.

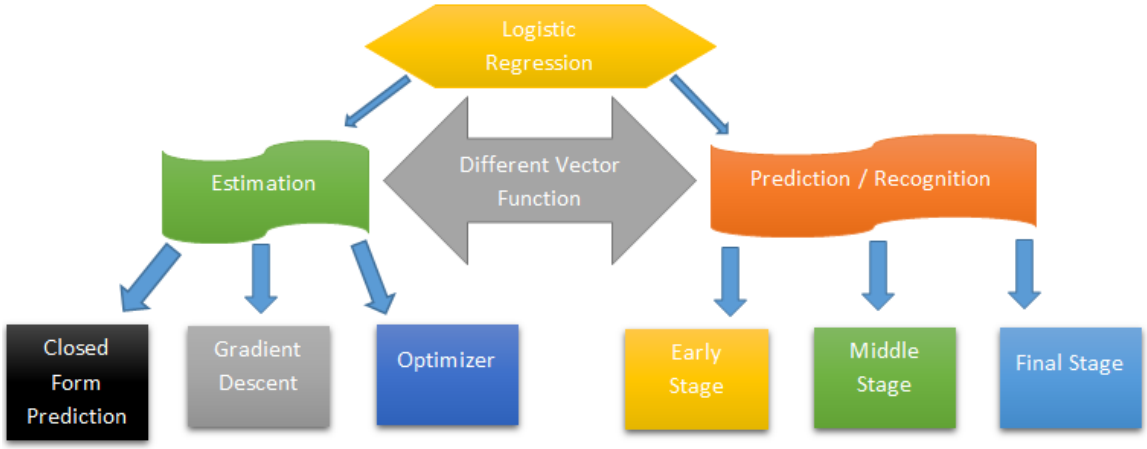

Figure 1. Review of logistic regression

1.1 Emerging Deep learning

Automatic detection of diabetic foot ulcer problems has also been achieved via the use of several sophisticated telemedicine monitoring systems [4,5]. Medical imaging techniques like visible and infrared light have been used in a variety of applications. Here, thermal imaging is a powerful tool. It's a temperature-based method for diagnosing foot problems. It's also a gentle approach that won't tear up your skin or otherwise hurt your body. Age, genetics, insulin levels, BMI, stress, pregnancy, and other factors may have a role. Some of these features may be included into the dataset as characteristics, and disease prediction research can be undertaken using the data. These factors are crucial to a person's well-being. However, the numerous cutting-edge machine learning systems available today may predict diabetic illness utilising individual features and combinations of these traits [6]. One way to prevent more complications from a sickness is to be aware of them in advance, which is what this article does.

The term "deep learning" is used to describe an area of machine learning that is predicated on the premise of level -based concepts. Understanding data like pictures, sounds, and text requires deep learning, which involves learning numerous layers of representation and abstraction. Artificial neural networks are the source of this deep knowledge, and a deep learning structure is a multilayer perceptron with additional hidden layers [7 -9]. The research aimed to create an advanced effective system for the early identification of diabetic foot utilising through risk factors. Creating a single classifier that performs well on all testing data is difficult in general [10, 11]. As a result, decision fusion may be an option. When a training model is only given a limited dataset from which to learn, decision fusion may help make the model broader and eliminate bias in the classification results.

1.2 Diabetes mellitus in pregnancy

During pregnancy, hormonal shifts and an increase in insulin production combine to cause this kind of disease.

1.3 Pre-Diabetes

The borderline diabetes is classified with the parameter- high blood sugar levels in fasting period.

Using past data as a foundation, machine learning may then anticipate future data. Logic is constructed from the taught data and tested on the test data, eliminating the need for programmers to write code. Experience-based prediction is a subfield in artificial intelligence.

1.4 Motivation

Motivated by the need to solve the diabetes prediction challenge by locating, critiquing, and combining the results for investigations, this paper undertook a study of the available diabetic prediction models. The goal in doing this review is to compile a comprehensive collection of literature on diabetes prediction models, and it is done so by including only those studies that meet the following criteria:

- The paper should have included a variety of prediction methodologies and machine learning algorithms for diabetes data categorization employing a wide range of highly optimised algorithms for early diagnosis.

- It's likely that the article should have explored several preprocessing approaches for filtering noisy data.

2. Literature Survey

The predicting diabetes in advance may help with its treatment, but many people with the disease go undiagnosed until their symptoms develop. As a result, the availability of a predictive tool for diagnosing diabetes is crucial. The artificial neural network, with the aid of the backpropagation neural network algorithm, is one of the few methods capable of making accurate predictions [12].

In order to better monitor and manage chronic conditions, scientists have focused heavily on creating AI-based tools and methodologies. In particular, many researchers have turned to ML models to estimate the likelihood of illness incidence given a set of potential risk variables. Computer-aided diagnostic techniques rely heavily on image segmentation and feature extraction. To improve CAD systems' precision, it is necessary to pick an appropriate segmentation approach and extract essential characteristics. Thermogram segmentation, for instance, was discussed in the research literature as a means of separating the plantar area from the backdrop [13].

Five auto-thresholding strategies were examined by Kaabouch et al. [14], including those based on histogram shapes, clustering, entropy, object attributes, and sophisticated evolutionary algorithms. Results were better when using genetic algorithms that used the thresholding method. There are many thermograms of healthy feet and some of the thermograms of feet affected by diabetes were subjected to a comparison of two segmentation techniques by Nandagopan et al. [15]. According to the findings, edge identification was more accurate than watershed analysis.

Some deep learning and classification strategies are covered by Sun and Zhang [16], For the purpose of diabetes data categorization, Jemmali et al., proposed the logistic regression, the authors improved classification accuracy to 92%. The main drawback of the model was that it could not be verified by being compared to other diabetes prediction models [17]. In their study, A.Farhadi et al. [18] split the dataset in half, setting aside half for training and the other for evaluation. Combining naive Bayes and support vector machine methods, a model was suggested for diabetes prediction. The suggested model was validated using data from all three of the original sources.

Lazy snapping has been shown to be an effective method for extracting the hottest/coldest areas from thermographic pictures, as proven by Etehadtavakol et al. [19]. Interactive picture algorithm "lazy snapping" separates processing into "large scale" and "fine scale," making it easy to modify object conditions and other finer points. Lazy snapping provides instant visual feedback by making it simple to distinguish between the split contour and the proper object border, even when the edges are blurry. The plantar area was subdivided by Saminathan et al. [20], using region-growing, and texture characteristics were recovered from 11 regions of the foot. The asymmetry of the study helped to distinguish between healthy tissue and ulcers. In order to extract the plantar area in the thermogram, Maldonado et al. [21], first utilised DL to segment the foot from the visible picture. They used temperature variations to categorise the danger of ulceration and necrosis.

3. Effective Early Detection of Diabetes

3.1 Diabetes Risk Prediction

The machine-learning models are form of crucial tool for doctors, clinicians, and health care professionals. Here, the potential for developing diabetes in the future is framed as a classification job with two possible outcomes:

C0 = "Diabetes"
C1 = "Non-Diabetes"

The input characteristics' values and, therefore, the probability of getting diabetes, may be used. The selected technique consists of four primary phases: data preparation, ranking attributes, training the classification model, and performance evaluation.

3.2 Data pre-processing

The proposed approach was evaluated and tested on the clinical dataset. There is a wide variety of illnesses included in the clinical dataset. In order to clean and extract characteristics from raw data, it must be converted into a data analysis file format. Medical treatment for diabetes is defined here in terms of the technique described in this page. The patient deviates from the picture of perfect health [22].

3.2.1 Data Cleaning

The data were collected in their raw form. No cleaning done in the dataset images.

3.2.2 Data Balancing

Prediction modelling is made more difficult by imbalanced classifications. Commonly used ML algorithm format for categories requires students to begin with a series of examples from each subject they are studying. False information is addressed and eliminated at this stage so that the final product is more thorough and accurate. Patient ID, age, diabetes pedigree function, smoking status, body mass index, insulin, skin thickness, blood pressure, gender, and glucose outcomes are all missing from this data collection.

Since these parameters must always have a value, we leave them blank. After normalising the data set, we found that all values were about the same. Since this study began, resampling methods have seen a dramatic development as a result of the findings. Most classifiers, for instance, may be concatenated, allowing for undersampling to be conducted by eliminating data from each category [23].

3.3 Improvements to prior efforts

The difficulty of using a prediction model is amplified when classifications are not uniformly distributed. Machine-learning algorithms that attempt classification often begin with a fixed number of examples for each class. As a consequence, stereotypes are reinforced, particularly among marginalised groups.

As the minority group grows in size and influence, it becomes more vulnerable to mistakes in categorization, which is a concern. Consequently, we have removed the extreme data points from the sample used in this study. Thanks to these studies, resampling methods have advanced considerably. Using record extraction from each cluster, for instance, we may do undersampling by combining the vast majority of class data. With oversampling, we may make slight adjustments to the data instead of creating perfect duplicates from the minority groups [24]. Figure 2 shows effective approach for early detection of diabetes.

3.3.1 Key factor for effective and early detection

In accordance with all systems, polydipsia and polyuria are classified first, followed by muscular stiffness, obesity, delayed healing, and itching. The relative rankings of the other characteristics are quite consistent among methods. Since all of the attributes are among the most frequent indicators of diabetes physicians look for, including the blood test for confirmation, they will all be utilized in the end to train and test the models [25].

3.3.2 Growth in capabilities

The goal of this strategy was to develop ML algorithm's potential use cases by leveraging data from a particular domain. It's the process of sifting through data sources and transforming raw information into machine-readable formats. A correlation matrix is used in the research to establish connections between the various datasets [25].

3.4 Cross-Validation

Re-evaluating models with a little amount of data is called cross-validation in machine learning domain. The number of data subsets to create from a given selection is controlled by a single parameter, k, which dictates the approach. This strategy goes under many other names, including k-fold cross-validation.

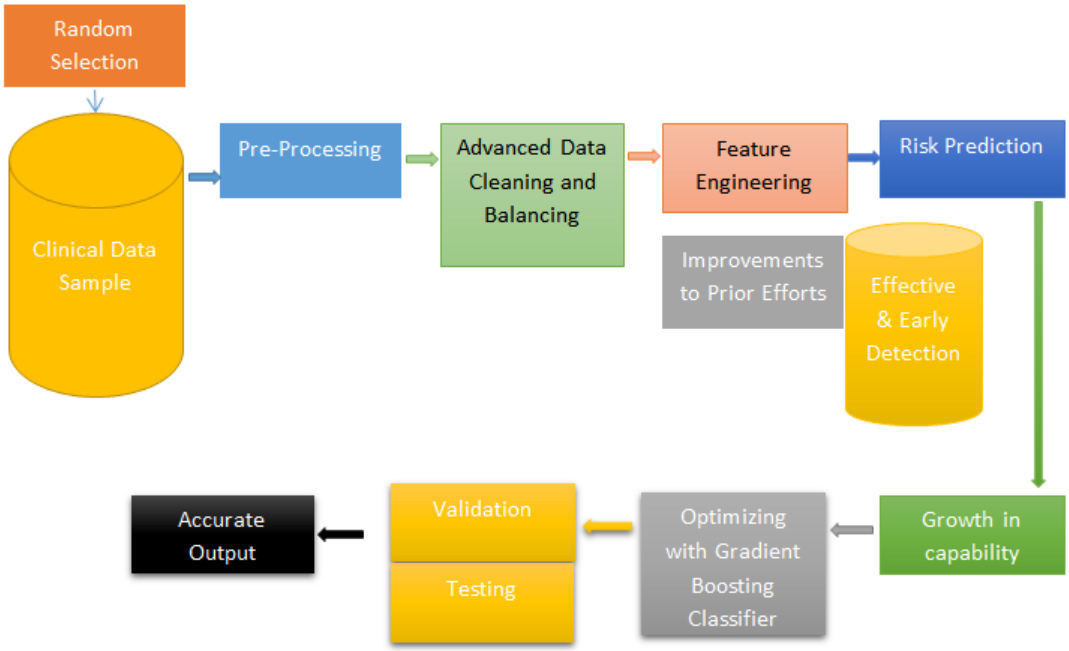

Figure 2. Effective approach for early detection of diabetes

3.5 Analysis of Variance using a K-Fold Cross-Validation

To test whether the model is valid for the last Kth fold, try fitting it with the previous K folds plus K minus 1.

3.6 Logistic Regression (LR)

For the purpose of classification, regression analysis is used. The dependent variable in logistic regression is often a yes/no question. In this case, a diagnosis of diabetes is either 1 or 0, respectively. An illustration of the model's logistical value is provided by the likelihood

of detecting a single test's various results. It assumes that the data and the variance in each acquired skill are Gaussian, which is not the case. Furthermore, this is an unobserved category. However, this classifier is better at identifying the specific illness. Logistic regression is only one example of many popular machine learning methods. Even if it's not complicated, it may be quite helpful in several scenarios.

3.7 Classifier using Gradient Boosting

This can integrate a large number of subpar learning models to provide a reliable prediction. Standard procedure calls for using decision trees to steepen the gradient. Gradient boosting is a machine learning approach for solving regression and classification problems by combining many, lower-quality models into one. Gradient trees, which are essentially decision trees with a weak learner, often provide better results than random forests. This method skips over the tedious step-by-step process of model construction and instead applies the model by minimising a loss function.

4. Future Direction

While just a subset of features has to be chosen for prediction, all existing approaches for diabetes prediction centre on feature selection strategies and a few of machine learning methods including;

- Random forest,
- Naive Bayes,
- Support vector machine,
- Decision trees.

These are some of the obstacles we experienced when reading all these articles:

- A big dataset was needed for prediction purposes, however the publicly accessible dataset for leads to individuals focusing on characteristics that are useless for making any kind of forecast.
- There are many challenges through missing vectors from the raw dataset and wrong interpretation in the reference dataset that may be future direction of this better prediction project.

Therefore, it is necessary to create an efficient optimizer technique that can provide better outcomes, handle data quickly, and make reliable predictions.

5. Conclusion

Diabetes is a lifelong, debilitating disease. Diabetes detection and treatment may be improved if performed at an earlier stage. This research also evaluates and contrasts many machine learning-based classification models for early diagnosis of diabetes in patients. Classifiers' efficacy was examined after the datasets were normalised. Predicting diabetic conditions ahead of time is the primary focus. The rising prevalence of diabetes is a direct consequence of contemporary human lives and practices. Thanks to the advances in machine learning, doctors can now evaluate patients' individual risks for developing diabetes and provide treatments and preventative measures accordingly. Potential links between the characteristics and diabetes may be uncovered by exploratory data analysis. The results of the performance study demonstrated the importance of data preparation in developing reliable and accurate models for the prediction of diabetes.

References

[1] D. U. N. Qomariah, H. Tjandrasa, and C. Fatichah, "Classification of diabetic retinopathy and normal retinal images using CNN and SVM," in Proceedings of the 2019 12th International Conference on Information & Communication Technology and System (ICTS), pp. 152–157, IEEE, Surabaya, Indonesia, 18 Jul. 2019.

[2] S. Sharma, "Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients," Environmental Science and Pollution Research, vol. 27, no. 29, pp. 37155–37163, 2020.

[3] Islam, M.M.F.; Ferdousi, R.; Rahman, S.; Bushra, H.Y. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis; Springer: Singapore, 2019; pp. 113–125.

[4] Saba, T. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. J. Infect. Public Health 2020, 13, 1274–1289.

[5] Hasan, M.K.; Alam, M.A.; Das, D.; Hossain, E.; Hasan, M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020, 8, 76516–76531.

[6] Kaur, H.; Kumari, V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Inform. 2020, 18, 90–100. [CrossRef]

[7] Kopitar, L.; Kocbek, P.; Cilar, L.; Sheikh, A.; Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci. Rep. 2020, 10, 11981.

[8] Tigga, N.P.; Garg, S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput. Sci. 2020, 167, 706–716.

[9] Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access 2021, 9, 103737–103757.

[10] Islam, M.; Ferdousi, R.; Rahman, S.; Bushra, H.Y. Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis; Springer: Berlin/Heidelberg, Germany, 2020; pp. 113–125.

[11] Alpan, et al G.S. Classification of diabetes dataset with data mining techniques by using WEKA approach. In Proceedings of the 2020 fourth International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 22–24 October 2020; pp. 1–7.

[12] Patel, S.; Patel, R.; Ganatra, N.; Patel, A. Predicting a risk of diabetes at early stage using machine learning approach. Turk. J. Comput. Math. Educ. (TURCOMAT) 2021, 12, 5277–5284.

[13] Elsadek, S.N.; Alshehri, L.S.; Alqhatani, R.A.; Algarni, Z.A.; Elbadry, L.O.; Alyahyan, E.A. Early Prediction of Diabetes Disease Based on Data Mining Techniques. In Proceedings of the International Conference on Computational Intelligence in Data Science, Chennai, India, 18–20 March 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 40–51.

[14] Kaabouch, N.; Chen, Y.; Hu, W.-C.; Anderson, J.W.; Ames, F.; Paulson, R. Enhancement of the asymmetry-based overlapping analysis rough features extraction. J. Electron. Imag. 2011, 20, 013012.

[15] Nandagopan, G.L.; Bhargavi, A.B. Implementation and Comparison of Two Image Segmentation Techniques on Thermal Foot Images and Detection of Ulceration Using Asymmetry Presented at ICCSP 2016. Available online: https://ieeexplore.ieee.org/abstract/document/7754155

[16] Y. Zhang, Z. Lin, Y. Kang, R. Ning, and Y. Meng, “A feedforward neural network model for the accurate prediction of diabetes mellitus,” International Journal of Scientifific and Technology Research, vol. 7, no. 8, pp. 151–155, 2018.

[17] Y. K. Qawqzeh, A. S. Bajahzar, M. Jemmali, M. M. Otoom, and A. -aljaoui, "Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling," BioMed Research International, vol. 2020, Article ID 3764653, 6 pages, 2020.

[18] A. Farhadi, D. Chen, R. McCoy, C. Scott, J. A. Miller, and M. Celine, N. Che, Breast cancer classification using deep transfer learning on structured healthcare data," in Proceeding of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 277–286, IEEE, Washington, DC, USA., October 2019.

[19] Etehadtavakol, M.; Emrani, Z.; Ng, E.Y.K. Rapid extraction of the hottest or coldest regions of medical thermographic images. Med. Biol. Eng. Comput. 2019, 57, 379–388.

[20] Saminathan, J.; Sasikala, M.; Narayanamurthy, V.B.; Rajesh, K.; Arvind, R. Computer aided detection of diabetic foot ulcer using asymmetry analysis of texture and temperature features. Infrared Phys. Technol. 2020, 105, 103219.

[21] Maldonado, H.; Bayareh, R.; Torres, I.A.; Vera, A.; Gutiérrez, J.; Leija, L. Automatic detection of risk zones in diabetic foot soles by processing thermographic images taken in an uncontrolled environment. Infrared Phys. Technol. 2020, 105, 103187.

[22] T. E. Idriss, A. Idri, I. Abnane, and Z. Bakkoury, "Predicting blood glucose using an LSTM neural network," in Proceeding of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 35–41, IEEE, Leipzig, Germany, September 2019.

[23] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia computer science, vol. 132, pp. 1578–1585, 2018.

[24] T. Anand, R. Pal, and S. K. Dubey, "Cluster analysis for diabetic retinopathy prediction using data mining techniques," International Journal of Business Information Systems, vol. 31, no. 3, pp. 372–390, 2019.

[25] T. Wang, W. Li, and D. Lewis, "Blood glucose forecasting using lstm variants under the context of open source artificial pancreas system," in Proceedings of the 53rd Hawaii International Conference on System Sciences, Maui, Hawaii, USA, January 2020.

Author's biography

K. Thangarajan is currently working as an Associate Professor in the Department of Electrical and Electronics Engineering, RVS College of Engineering and Technology, Coimbatore, India. His research includes Electric drives, control systems and AI techniques.