International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Prediction and Management of Diabetes using Machine Learning: A Review

Nair Ul Islam[1], Syed Imtiyaz Hassan[2],
[1, 2]*Department of CSE, Jamia Hamdard University, New Delhi, India.*

Abstract*: Diabetes Mellitus is major health concern of today's day and age. It affects people of all ages, all around the world. So, far no cure has been found for diabetes. However, it can be managed by medicines, maintaining a healthy lifestyle and medications. It can be a source of considerable burden for people affected by it both psychologically and financially. So, early detection of diabetes becomes necessary as it will help a patient in the longer run. Machine Learning has lately been used in health industry for the prediction and management of diseases. This study focuses on diabetes so, only those systems will be studied here that focus on diabetes. These systems were designed with a goal in mind of predicting or managing diabetes. It was found out that these algorithms used a wide range of algorithms and had a good performance. The algorithm that was used in multiple of those systems and perhaps the most popular one was Support Vector Machine. It was found out that it had high accuracy on a consistent basis.

Keywords: Machine Learning, Diabetes, Predictive Models, Health care, SVM

I. INTRODUCTION

Diabetes Mellitus is a disease which is a source of burden on the patients as well as on the health care systems around the world [1]. It is a growing disease and it is estimated by the year 2035 more than 592 million people all around the world will be affected by diabetes [1]. It occurs when the pancreas stops making insulin or when the cells within the body stop responding to insulin [2]. No cure is available for diabetes as of now but it is possible to delay or control it with proper diet [3]. Diabetes puts the patient at risk of various other diseases like nerve damage, kidney disorders, blindness and heart disease [4]. The percentage of diabetic people who die from a heart disease or some blood vessel disease stands at higher than 80% [4]. Some of the people who acquire diabetes are not aware about the high-risk diseases associated with it [5]. Diabetes Mellitus is a heterogeneous group of genetic and clinical disorders, all sharing glucose intolerance [6]. Diabetes can mainly be classified into groups, Type 1 and Type 2; Type 1 is mostly found in juveniles and it has a sudden abrupt of symptoms while as Type 2 is associated with least or sometimes no symptoms [6]. Computer science and Machine learning has made long strides and is now being used in medical settings [7] for the better diagnosis and treatment of patients.

Machine learning techniques can make use of EHR (electronic health record) to predict the susceptibility of a disease in a person [8]. The data associated with each patient encounter is stored in an EHR [9, 10]. Along with numerous other challenges in the medical settings; machine learning has been used to tackle diabetes as well. Many kinds of models, approaches and systems have been proposed. In this study an attempt has been made to review some of those approaches which have been proposed.

II. LITERATURE REVIEW

For the medical diagnosis of diabetes, a model [11] was proposed. It was more of a hybrid model as it integrated in itself a supervised machine learning algorithm along with an unsupervised machine learning algorithm. This model made use of a dataset based on the real world and it outperformed some of the other systems which were working on similar problems. SVM was the algorithm used for the forecasting of diabetes along with a rule-based component. The intention behind this model was to build a system that could be used as a second opinion for the forecasting of diabetes in population at high risk. The model performed well and had a high prediction accuracy.

For the classification of persons to determine whether a person has a disease or not, an approach [12] making use of SVM was proposed and for purposes of illustrations diabetes was selected as the test case. This study gave promising results and helped find out if a person has diabetes or not. This experiment had a high success rate.

A study [13] was carried out to find the effectiveness of machine learning algorithms in predicting diabetes. The study was conducted based on a set of ten parameters which may increase the chances of developing diabetes in an individual. The algorithms used for this study were Naive Bayes Classifier, KNN and ANNs. Upon the completion of this study it was found that ANNs had the best performance in predicting accuracy of the outcome followed by Naive Bayes and KNN; their accuracies were 96%, 95%

and 91% respectively. The researchers argued that the performance can be made even better if the training data size were increased and various other factors incorporated and identified.

A mobile phone app [2] was developed to forecast the occurrence of diabetes in a person. This app would present the user with a questionnaire that was to be filled. It was filled with questions related to the diabetic symptoms, activities performed by the user and questions related to the user, like height weight etc. The questionnaire was compiled by medical experts. The user would fill the form, these inputs then were fed to the machine learning back end of the mobile application for training purposes. A real-world diabetes dataset was used for evaluating the performance of the machine learning models. Four machine learning algorithms used were Multi-level perceptron, Naive Bayes, Polykernal SVM and J48. It was found out that J48 gave the highest results on sensitivity, specificity and ROC of 0.890, 0.928 and 0.928 respectively.

A cascade learning system model [4] was put forth in order to classify the diabetes. The algorithms used in this study were Least square support vector machine and (LS-SVM) and Generalized Discriminant Analysis (GDA). This system consisted of two stages, stage 1 and stage 2. The first stage made use of GDA with an aim to discriminant feature variables between diabetic patients and healthy persons. This stage is a part of data pre-processing. Now, the second stage making use of the LS-SVM is what was used for diabetic classification. This stage had an accuracy of 82.05% for classification. The researchers concluded this study with the statement that this method can be used in the assistance for the detection of diabetes.

A study [14] was carried out to find out the relationship between HW phenotype and type 2 diabetes, using machine learning. A total of 11937 subjects were chosen for this study out of which 4806 were men and 7031 females. Naive Bayes and Logistic Regression were the algorithms used in this study. This study showed a strong evidence between the HW phenotype and type 2 diabetes and it showed the association was much more pronounced in men than in women.

A system [15] using Artificial Neural Network and Fuzzy Neural network was developed for classifying heart disease and diabetes. Dataset used in this study was from the machine learning repository of the University of California, Irvine. The researchers claimed that the hybrid model they had proposed based on the Artificial Neural Network and Fuzzy Neural Network had one of the best accuracies when compared to similar systems existing at that time.

Another study [16] was carried to determine the risk of incident related diabetes. This study made use of the Henry Ford FIT dataset. This dataset contained a recorded of 32,555 patients who did not have any heart or canary artery disease. These patients were had to go through a clinician referred exercise treadmill stress and after five years it was found out that the number of those patients who had acquired diabetes after a period of five years stood at 5099. Naive Bayes was the best performing algorithm with the precision rate of 86.7%. It was closely followed by Random Forest and Logistic Regression. Random Forest had an accuracy of 84.3% and Logistic Regression had an F1-score of 91.5%.

Use of Support Vector Machine for the diagnosis was proposed in a paper [17]. The key idea here was to present a intelligible representation of the outcome that the Support Vector Machine had achieved. The system proposed here was based on a hybrid model; for model building and sampled a combination of a supervised and unsupervised algorithms was used. Also, a rule-based explanation component was added to it. The output of the predictions made by this model was supposed to be used as a second opinion in the prediction of the diabetes. The researchers claimed that their system was very accurate as it had a high accuracy rate.

For the classification of diabetes, a model [18] was proposed. It made use of the Support Vector Machine and a high dimensional dataset obtained from the machine learning repository of University of California, Irvine. It was argued that the most optimal values for the parameters for a particular kernel is critical and related to the amount of data that is available. The accuracy, sensitivity and specificity metrics for this model were 78%, 80% and 76.6% respectively.

Approaches [19] towards the use of machine learning techniques on electronic health records were proposed with a goal in mind to get a valuable insight about the disease processes. The model that was built here was a predictive model for progression from pre diabetes to post diabetes. The platform on which this model was based on is Reverse Engineering and Forward Simulation (REFS). REFS explored a wide model space, relying on Bayesian scoring algorithm. The output of REFS is the distribution of risk estimation obtained from ensemble of a range of prediction models. It predicted the progression of type 2 diabetes accurately with an AUC of 0.76. This was a hypotheses free analytical approach for checking the progression of diabetes and it made very accurate predictions.

Machine Learning algorithms and SMBG values were used to train a model [20] for the prediction of hypoglycaemia in patients with type 2 diabetes. A number of datasets were used for the validation of the model. 10 week SMBG value was the optimal number required by the model. 92% sensitivity and 70% was the specificity was recorded the model. This model had a high level of specificity and sensitivity when predicting hypoglycaemia.

A model [21] that automatically warned of changes in the blood glucose level was proposed. It made use of the support vector regression that was trained on data specific to patients and a physiological model of blood glucose dynamics that was set to generate informative features for the support vector regression. Diabetic experts were outperformed by the predictions made by this model and it could predict hypoglycaemic events 30 minutes before they occurred. The precision of this model stood low at 42%.

A system [22] was introduced for the automatic diagnosis of diabetes. It was based on Morlet Wavelet Support Vector Machine Classifier and Linear Discriminant Analysis (LDA). This system was made up of three stages, LDA was used for feature classification and reduction, and a classification stage that made use of the Morlet Wavelet Support Vector Machine Classifier. Features of the healthy patients were obtained in the first stage which were then fed to the second stage. In the third and final stage the correctness of the diagnosis made by the first two stages was done using the metrics of specificity, sensitivity, confusion matrix and classification accuracy. 89.74% was the classification accuracy of this system.

III.CONCLUSION

In this review paper many models and approaches based on machine learning techniques aimed at predicting and managing diabetes were discussed. Many more equally good approaches and models do exist that have not been discussed. We live now in a time where gigantic quantities of data are generated every minute. Information is considered to be the most valuable asset of this day and time. Earlier the medical settings didn't maintain the large records of the patients but with the advent of information age even the medical settings have been maintaining records of the patients digitally, popularly known as the electronic health records. Today, we have access to many of those EHRs. This data available to us can be leveraged to make machine learning models that predict the occurrence of a disease in a person, help us managing it or even help us in checking its progression. Diabetes was the disease that was focused upon in this study. Many approaches and models based on machine learning geared towards tackling diabetes were studied. It was observed that these models were successful at predicting its occurrence with high accuracy. Some of those models were geared toward managing diabetes or checking its progression and they performed good as well. The algorithms that most frequently occurred in most of these models with consistent high accuracy rates for prediction were Support Vector Machine and Naive Bayes. We can conclude this study with the statement that most of these systems perform well in the detection of the diabetes and can be used as a second opinion..

REFERENCES

[1] Donsa, K., Spat, S., Beck, P., Pieber, T.R. and Holzinger, A., 2015. Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges. In Smart Health (pp. 237-260). Springer, Cham.
[2] Sowjanya, K., Singhal, A. and Choudhary, C., 2015, June. MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 397-402). IEEE.
[3] Çalişir, D. and Doğantekin, E., 2011. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), pp.8311-8315.
[4] Polat, K., Güneş, S. and Arslan, A., 2008. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert systems with applications, 34(1), pp.482-487.
[5] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S., 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one, 12(7), p.e0179805.
[6] National Diabetes Data Group, 1979. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. Diabetes, 28(12), pp.1039-1057.
[7] Firdaus, H., Hassan, S.I. and Kaur, H., 2018. A Comparative Survey of Machine Learning and Meta-Heuristic Optimization Algorithms for Sustainable and Smart Healthcare. AFRICAN JOURNAL OF COMPUTING & ICT, p.1.
[8] Wu, J., Roy, J. and Stewart, W.F., 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care, pp.S106-S113.
[9] Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P., 2018. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), pp.1589-1604.
[10] Birkhead, G.S., Klompas, M. and Shah, N.R., 2015. Uses of electronic health records for public health surveillance to advance public health. Annual review of public health, 36, pp.345-359.
[11] Barakat, N., Bradley, A.P. and Barakat, M.N.H., 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE transactions on information technology in biomedicine, 14(4), pp.1114-1120.
[12] Yu, W., Liu, T., Valdez, R., Gwinn, M. and Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), p.16.
[13] Sarwar, A. and Sharma, V., 2014. Comparative analysis of machine learning techniques in prognosis of type II diabetes. AI & society, 29(1), pp.123-129.
[14] Lee, B.J. and 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), pp.39-46.
[15] Kahramanli, H. and Allahverdi, N., 2008. Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), pp.82-89.

[16] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S., 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one, 12(7), p.e0179805.

[17] Barakat, N., Bradley, A.P. and Barakat, M.N.H., 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE transactions on information technology in biomedicine, 14(4), pp.1114-1120.

[18] Kumari, V.A. and Chitra, R., 2013. Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), pp.1797-1801.

[19] Anderson, J.P., Parikh, J.R., Shenfeld, D.K., Ivanov, V., Marks, C., Church, B.W., Laramie, J.M., Mardekian, J., Piper, B.A., Willke, R.J. and Rublee, D.A., 2016. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. Journal of diabetes science and technology, 10(1), pp.6-18.

[20] Sudharsan, B., Peeples, M. and Shomali, M. (2014). Hypoglycemia Prediction Using Machine Learning Models for Patients With Type 2 Diabetes. Journal of Diabetes Science and Technology, 9(1), pp.86-90.

[21] Plis, K., Bunescu, R., Marling, C., Shubrook, J. and Schwartz, F., 2014, June. A machine learning approach to predicting blood glucose levels for diabetes management. In Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.

[22] Çalişir, D. and Doğantekin, E., 2011. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), pp.8311-8315.