Paper ID: 322

Diabetes Prediction System using Machine Learning

TK Thenabadu# and WMKS Ilmini

Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

#thamindu@gmail.com

Abstract: Diabetes is a deadly chronic disease which affects entire body system harmfully. Millions of people are affected by this disease and a considerable number of patients die every year because of its side effects. A diabetic patient suffers from a high level of blood sugar in the body. Undiagnosed diabetes may cause the nerve and kidney damage, heart and blood vessel disease, slow healing of wounds, hearing impairment and several skin diseases. Early detection of diabetes is very essential to have a healthy life. The recent development of Machine Learning approaches solves this kind of critical problems. The main objective of this study is to present a Machine Learning based solution (Artificial Neural Network) to solve the above problem. And also, the technologies and approaches used in previous researches to predict diabetes have been reviewed with their accuracy levels. All the previous studies have used "Pima Indian Diabetes Dataset" (PIDD) as the dataset but this research is based on a newly collected dataset. The overall development process can be categorized into four major development phases namely data collection and preprocessing, statistical analysis, development of machine learning model and development of front-end. Artificial Neural Network model has been developed and deployed while the model provides more than 92% accuracy on the sample testing dataset.

Keywords: Diabetes, Machine Learning, Artificial Intelligence, Artificial Neural Network, Android, Tensorflow, Firebase

Introduction

Diabetes is considered as one of the deadliest chronic diseases which can affect the entire body system adversely and there are millions of people affected by diabetes all over the world. According to WHO (Who.int, 2019), there were 422 million diabetes patients and it is 8.5% of the total population in the world. Diabetes increase the nerve and kidney damage, heart and blood vessel disease, slow healing of wounds, hearing impairment and also several skin diseases. In 2016, 1.6 million deaths were recorded because of diabetes. Almost half of all deaths occurred before the age of 70. According to WHO, diabetes was the 7th leading cause of death in 2016.

According to the research done by Katulanda and others (Katulanda et al., 2009), it is mentioned that by 2025 South East Asia will be the region with the highest number of diabetes patients in the world. Early identification of diabetes is important to have a healthy life. A healthy diet, regular exercises, maintaining average body weight, and eliminating alcohol and smoking are ways to prevent or delay the onset of type 2 diabetes. Diabetes can be controlled, and its consequences avoided or delayed with diet, physical exercises, medication and treatment for complications.

Machine Learning is an Artificial Intelligence application which provides systems with the ability to learn and improve from its own experience without being explicitly programmed. Previous studies which have been done to predict diabetes used some machine learning classification algorithms

such as Support Vector Machine (SVM), Decision Tree, Random Forest (RF), Naïve Bayes and Neural Network. All most all the previous studies have used "Pima Indian Diabetes Dataset" (PIDD) as the dataset. The performance of those various Machine Learning models has been reviewed in terms of accuracy and precision.

A new dataset has been collected to build the machine learning model and the related risk factors were originally discovered. Discovered risk factors are gender, age, BMI, waist circumference, frequency of doing exercises and having fruits/vegetables,high blood pressure, and risk score. This research aims to improve health care in Sri Lanka through, predicting diabetes at an early stage and giving recommendations to maintain a healthy life.

Literature Review

According to the World Health Organization, (Who.int, 2019) diabetes has become a leading cause of death in the world. Most of the diabetes patients are in low- and middleincome countries. A lot of researches have been done specifically using Machine Learning and Neural Networks in diagnosing diabetes mellitus and some approaches are discussed including their aim, materials and methods used, results and conclusion.

Kaur and Kumari (Kaur and Kumari, 2018) have discussed "Predictive modelling and Analytics for Diabetes using Machine Learning" in their research paper. The main aim of that research was to find out what is the most accurate predictive model to predict diabetes mellitus among 5 predictive models which are known as "Linear Kernel" and "Radial Basis Function" (RBF), "Multifactor Dimensionality Reduction" (MDR), "k-Nearest Neighbor" (kNN), "Kernel Support Vector Machine" (SVM) and "Artificial Neural Network" (ANN). They have used R data manipulation tool to investigate the diabetes dataset, which is known as Pima Indian Diabetes Dataset, originally owned by the "National institute of diabetes and digestive and kidney diseases", India. This dataset contains 768 instances classified into two classes; diabetic and nondiabetic. And there are eight different risk factors. They have trained their model with 70% training data and tested with 30% remaining data. Those five different models developed using supervised learning methods mentioned above have been experimented in R programming studio.

Table 1: Accuracy of different Predictive models. (Kaur and Kumari, 2018)

No	Predictive Model	Accuracy
1	Linear Kernel SVM	0.89
2	Radial Basis Kernel SVM	0.84
3	k-NN	0.88
4	ANN	0.86
5	MDR	0.83

According to the above table, Linear Kernel SVM model is the most accurate model among those 5 predictive models.

From the above research, it can be said that "SVM-linear" and "k-NN" are the two best models to predict diabetes.

Zou and others (Zou et al., 2018) have discussed the "Prediction of Diabetes Mellitus with Machine Learning Techniques in their research paper". The main objective of this study was to find out which is the most accurate machine learning technique to predict diabetes mellitus. Researchers have obtained two different datasets named as Luzhou dataset and Pima dataset. Luzhou dataset was obtained by the hospital physical examination data in Luzhou, China. And there are two parts of this dataset; healthy people and the diabetic people and it contains 14 different examination indexes. And the Pima dataset was the same as the previous research mentioned above. "Decision Tree" (DT), "Random Forest" (RF) and "Neural Networks" have been used as the

classifiers. DT and RF were implemented in WEKA while Neural Network was implemented in MATLAB. In this study, J48 decision tree was used in WEKA.

Table 2: Accuracy of different classifiers in Luzhou dataset. (Zou et al., 2018)

No	Classifier	Accuracy
1	Decision Tree (J48)	0.7853
2	Random Forest	0.8084
3	Neural Network	0.7841

Table 3 : Accuracy of different classifiers in Pima dataset. (Zou et al., 2018)

No	Classifier	Accuracy
1	Decision Tree (J48)	0.7275
2	Random Forest	0.7604
3	Neural Network	0.7667

By comparing the results of three classifications, there is not much difference among the three classifications, but random forest is better than other 2 classification methods. The value 0.8084 is the best accuracy in the Luzhou dataset while 0.7667 is the best accuracy in the Pima Indian dataset. And those results proved that machine learning can be used to predict diabetes.

Deepti Sisodia and Dilip Sisodia (Sisodia and Sisodia, 2018) have discussed the prediction of diabetes using Classification Algorithms namely Naïve Bayes, Decision Tree ans SVM. The main objective of this study is to design a model which can predict the possibility of diabetes with maximum accuracy. Pima Indian Diabetes dataset used as the main dataset and WEKA Tool was used for data classification.

According to the above table, Naive Bayes classification is the most accurate classification comparatively other two algorithms.

Table 4 : Accuracy of different classifiers. (Sisodia and Sisodia, 2018)

No	Classifier	Accuracy
1	Naive Bayes	0.7630
2	SVM	0.6510
3	Decision Tree	0.7382

Pradhan and Sahu (Pradhan and Sahu, 2011) have observed using Artificial Neural Networks to predict diabetes. The main objective of this study was to find out what is the most accurate classification model to predict diabetes mellitus among 7 classification models which are known as FLANN (Functional Link Artificial Neural Network), Novel Artificial Neural Network, MFS1 (multiple feature subset), MFS2 (multiple feature subset), Nearest neighbors with NN, KNN, BSS. And Genetic Algorithm is used for feature selection and trained with Back Propagation algorithm.

Table 5: Accuracy of different classifiers. (Pradhan and Sahu, 2011)

No	Classifier	Accuracy
1	NN	0.651
2	kNN	0.697
3	BSS	0.677
4	MFS1	0.685
5	MFS2	0.705
6	Novel ANN	0.734
7	FLANN	0.598

According to the above table, it is revealed that their suggested Novel ANN is performing better compared to other 6 classification algorithms.

Saru and Subashree (Saru and Subashree, 2019) have discussed analyzing and predicting diabetes using Machine Learning. The main aim of this research was to find out which is the most accurate machine learning technique to predict diabetes mellitus among the selected classifiers. Pima Indian Diabetes Dataset was used as the dataset. Logistic regression with SVM, Decision tree(J48),

kNN(k=1) and k-NN(k=3) are the classifiers. According to the results, Decision tree(J48) has the highest accuracy of 0.944.

Islam and Jahan (Aminul and Jahan, 2017) also discussed various Machine Learning methods which can be used in diabetes prediction. The main aim of this research was to find out which is the most accurate machine learning technique to predict onset diabetes among the selected classifiers. Naïve Bayes (NB), Logistic Regression (LR), Multilayer perception (ML), Support Vector Machine (SVM), IBK, AdaBoostM1, Bagging, OneR, J48 and Random Forrest are the selected classifiers. Pima Indian Diabetes Dataset was used to train the models. According to the results, Logistic Regression performed the best accuracy among all 10 classifiers.

Sneha and Gangil (Sneha and Gangil, 2019) have researched the optimal features that can be used for early prediction of diabetes mellitus. The main aim of this study is to design a model which can predict the possibility of diabetes with maximum accuracy. SVM, Random Forest, NB, Decision Tree and KNN are the 5 classification algorithms used in this study and the used dataset was Pima Indian Diabetes dataset.

No	Predictive Model	Accuracy
1	SVM	0.7773
2	Random Forest	0.7539
3	NB	0.7348
4	Decision Tree	0.7318
5	KNN	0.6304

Table 6 : Accuracy of different classifiers. (Sneha and Gangil, 2019)

According to the above table, SVM has the highest accuracy compared to other 4 algorithms.

Table 7: Comparison of the best models of the papers

Paper	Best Model	Accuracy
Kaur and Kumari, 2018	Linear Kernel SVM	0.8900
Zou et al., 2018	Random Forest	0.8084
Sisodia and Sisodia, 2018	Naïve Bayes	0.7630
Pradhan and Sahu, 2011	Novel ANN	0.7340
Saru and Subashree, 2019	Decision Tree	0.9440
Aminul and Jahan, 2017	Logistic Regression	-
Sneha and Gangil, 2019	SVM	0.7773

All the previous researches mentioned in Table 7 is based on "Pima Indian Dataset". Only the Machine Learning Methodologies and Models that have been selected are different from one research to another. Because of that, it is fair to compare the best models in each previous research to identify the best Machine Learning Model. According to the above table, it is clear that the Decision Tree in the 5th paper has the highest accuracy. But each model has its advantages, as well as disadvantages. And the accuracy level depends on the methodology which they have used.

Methodology

A. Anatomy of Diabetes Predicton System

This section discusses the design and the implementation of the Diabetes Prediction System which is being developed using Artificial Neural Networks. The system aims to improve the healthcare system in Sri Lanka through, predicting diabetes in an early stage and giving recommendations to maintain a healthy life. So, the most important function of the system should be the prediction part. The basic procedure of this system contains 3 major steps. First, the

user should enter some details through the mobile application. Then the backend of the system provides an accurate prediction based on the previously trained model. Finally, the system provides recommendations to the user to stay healthy. Figure 1 shows the overall solution for the research problem.

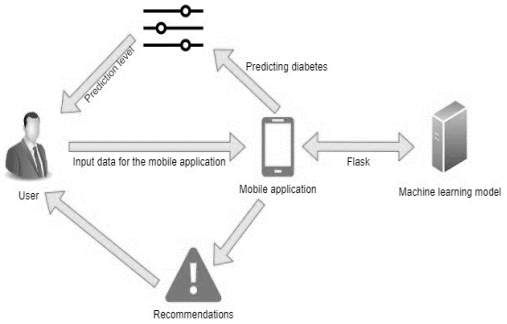

Figure 1: Overall solution diagram

The overall architecture of the system can be depicted as sub-systems such as problem identification and algorithm selection, data collection and training, prediction and finally the mobile application integration. Figure 2 shows the high-level architecture of the Diabetes Prediction System.

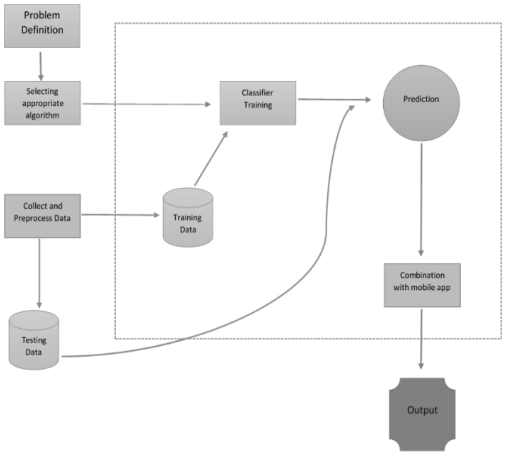

Figure 2: High-level architecture

The following use case diagram represents how the user interacts with the diabetes prediction system. The core of the system is the ANN model. The user can access the trained model through the Android application. The Android application should be able to predict the results according to the user inputs.

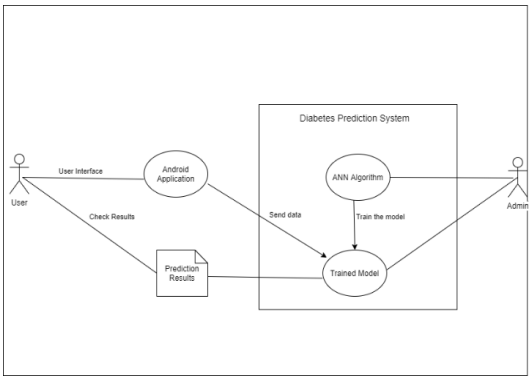

Figure 3: Use case diagram of the prediction system

B. Overall Development Process

The overall development process can be categorized into four major development phases namely data collection and preprocessing, statistical analysis, development of machine learning model and development of front-end. A brief description of each development phase is given below.

1. Data collection and preprocessing

A sufficient amount of data needs to feed the Machine Learning algorithm to have good accuracy. Pima Indian Dataset, which has been used by previous research works, contains only female data which may not be able to predict diabetes of males. Because of that, it is decided to collect a dataset which is suitable for both male and female to predict diabetes and also the related risk factors have been originally discovered. Related risk factors were identified using the research done by Lindstrom and Tuomilehto related to a diabetes risk score of type 2 diabetes. (Lindstrom and Tuomilehto, 2003)

Google forms and printed surveys are used to collect the dataset. The collected dataset contains both male and female data of different age groups to balance the dataset

2. Statistical analysis

A correlation analysis has been done to check whether the selected risk factors are suitable for predicting Diabetes. According to Kumar and Chong "Correlation analysis is an

extensively used technique that identifies interesting relationships in data. These relationships help us realize the relevance of attributes concerning the target class to be predicted."(Kumar and Chong, 2018). So, it is very important to do a correlation analysis to check whether the selected factors are connected in a way that can produce good accuracy in the model. R software, which is very useful in statistical analysis has been used to do the correlation analysis (Lafaye de Micheaux et al., 2013). Results of the statistical analysis are discussed under the results section.

Development of Machine Learning model

In this stage, the Artificial Neural Network based Machine Learning algorithm has used for predicting diabetes. Libraries like Tensorflow and Pandas are mostly used. After preprocessing the collected dataset, training and testing the model should be done. The generated model should be converted into a TensorFlow Lite model which can be implemented in the Android application. (TensorFlow Lite | ML for Mobile and Edge Devices, 2020)

ANN uses a sequential model which contains three dense layers. In the first layer there are 12 nodes and the activation function is ReLu. In the second layer there are 8 nodes and the activation function is ReLu. In the third layer there is only 1 node and the activation function is Sigmoid.

Development of Front-End

Since this is an android based research, the front-end is being developed using Android Studio. The authentication part has been done using Google Firebase (Khawas and Shah, 2018). Machine learning model conversion has been done using TensorFlow Lite. Main modules of the Android application would be Diabetes Info Center, Prediction module, Recommendation module, Diabetic Diet and User profile.

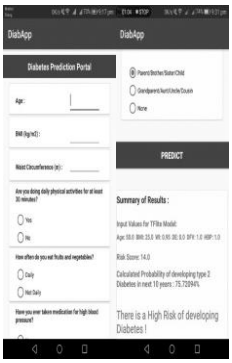

Figure 4: Prediction module user interfaces

Results

Statistical Analysis

Selected risk factors used in the statistical analysis are mentioned with their abbreviations below.

- GND – Gender
- AGE – Age
- BMI – Body Mass Index
- WC – Waist Circumference
- DE – Daily physical activities for at least 30 minutes? (including normal daily activities)
- FVD – How often do you eat fruits and vegetables?
- HBP – Have you ever taken medication for high blood pressure?
- HBG – Have you ever been found to have high blood glucose?
- RS – Risk Score
- DP – Diabetes Patient or not

A sample dataset has been plotted using R software and the correlation was calculated from 3 different methods. (Spearman Method, Kendall Method, Kendall Method). Related graphs for correlation analysis process is given below.

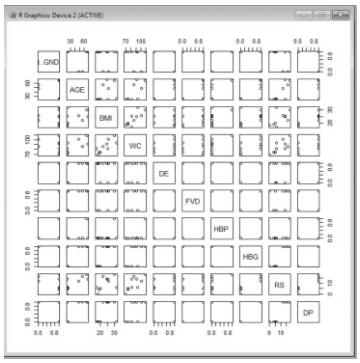

Figure 5: Plot of Risk Factors

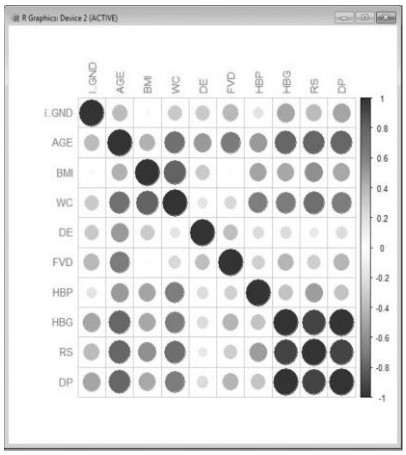

Figure 6: Correlation Analysis (Graphical)

According to the above diagrams, there are several strong correlations between risk factors. Because of that, the selected risk factors are suitable for use in the machine learning model.

B. Machine Learning model

Results related to machine learning model can be categorized into two sub parts, accuracy on the training dataset and accuracy on the testing dataset which the model hasn't been seen before. Those accuracies calculated from collected dataset are mentioned below.

Accuracy on the training dataset = 93.65%

Accuracy on the testing dataset = 92.48%

According to the above graph, the ANN model has achieved 92.48% accuracy on the testing dataseset. Pradhan and Sahu (Pradhan and Sahu, 2011) have observed using Artificial Neural Networks to predict diabetes but the achieved accuracy was 73.40%. Other researches discussed under literature review, used different machine learinng approaches to predict diabetes. (Table 7)

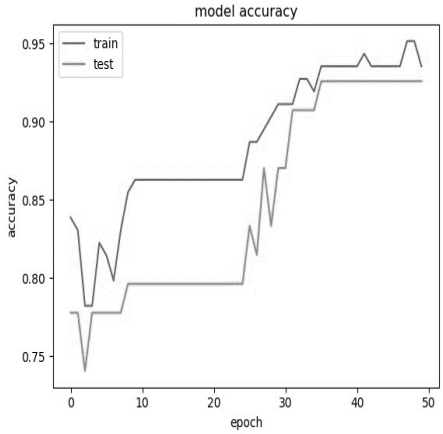

Figure 7: Training and Validation graph

Discussion and Conclusion

Diabetes has no cure, but early detection of Diabetes can reduce severe health issues and the cost for medicine. Because of that, it is very important to predict it as soon as possible. This paper has reported our research on the use of an ANN in predicting type two diabetes. Several researches have been done research works to predict diabetes using Machine Learning techniques. "Pima Indian Diabetes Dataset" only contains 298 instances and it is not enough to build a strong predictive model. And, the dataset contains instances only about females. Because of that, the predictivity of diabetes for a male is not so accurate.Since this research has been done to predict diabetes of both males and females, instead of using PIMA Indian dataset, a new dataset has been collected. The overall development process can be categorized into 4 major development phases namely data collection and preprocessing, data analysis, development of machine learning model and development of front-end. By observing accuracy values of the machine learning model, it is clear that the developed ANN is usitable of using in predicting diabetes of both males and females.

Future Works

The research has not been completed yet. Only the data collection and machine learning model has been implemented in the Android environment. Prediction Module has implemented in the Android application. Features like recommendation system will be added to the Android application in the future. preprocessing, statistical analysis, development of the machine learning model have been completed.

References

Aminul, Md., Jahan, N., 2017. Prediction of Onset Diabetes using Machine Learning Techniques. Int. J. Comput. Appl. 180, 7–11.

Diabetes [WWW Document],n.d. URL https://www.who.int /westernpacific/health-topics/diabetes (accessed 10.19.19).

Katulanda, P., Sheriff, M., Matthews, D., 2009. The diabetes epidemic in Sri Lanka – a growing problem. Ceylon Med. J. 51, 26.

Kaur, H., Kumari, V., 2018. Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Inform. S221083271830365X.

Khawas, C., Shah, P., 2018. Application of Firebase in Android App Development-A Study. Int. J. Comput. Appl. 179, 49–53.

Kumar, S., Chong, I., 2018. Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. Int. J. Environ. Res. Public. Health 15, 2907.

Lafaye de Micheaux, P., Drouilhet, R., Liquet, B., 2013. The R Software: Fundamentals of Programming and Statistical Analysis, Statistics and Computing. Springer New York, New York, NY.

Lindstrom, J., Tuomilehto, J., 2003. The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. Diabetes Care 26, 725–731.

Pradhan, M., Sahu, D.R.K., 2011. Predict the onset of diabetes disease using Artificial Neural Network (ANN).Emerg. Technol. 2, 9.

Saru, S., Subashree, S., 2019. ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING 5, 9.

Sisodia, D., Sisodia, D.S., 2018. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 132, 1578–1585.

Sneha, N., Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. J. Big Data 6, 13.

TensorFlow Lite | ML for Mobile and Edge Devices [WWW Document], n.d. . TensorFlow. URL https://www.tensorflow.org/lite (accessed 7.24.20).

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H., 2018. Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9, 515.