International Journal of Research in Engineering, Science and Management
Volume 4, Issue 11, November 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

A Survey On Diabetes Prediction Using Machine Learning Techniques

A. Rajivkannan[1], K. S. Aparna[2*]
[1]Professor, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, India
[2]M.E Student, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, India

***Abstract*: In the recent years, most of the people in the world are affected with blindness, kidney failure, amputations, heart failure and stroke. It is found that all of these are caused by most probably by diabetes disease. While food eaten is being converted into sugars or glucose, the pancreas releases insulin, which servers as an important key to open the cells of the human body and it allows the glucose to enter and it allows to use the glucose for energy. If the insulin is not properly created or released, several things may go wrong, which in turn onset of diabetes. This disease is classified as Type 1, Type 2 which are commonly occurred and Gestational diabetes occurred during pregnancy. To detect and predict this disease in the earlier stage will help the people to get relief from it. Different machine learning approaches have been proposed to predict the diabetes using data sets like PIMA dataset. This paper includes the comparative study of performance metrics for various machine learning approaches have been used.**

***Keywords*: Accuracy, Detection, Diabetes, Machine Learning, Prediction.**

1. Introduction

The It is essential to know how the glucose level may increase in the human body. When we eat the carbohydrate foods, it is to be broken into glucose by the body. Our brain consumes some of the glucose and remaining is used by cells of our body for energy and also to our liver. The harmone 'Insulin' attached to the cell is acting as 'doors' which allows glucose to enter into the cell. If pancreas does not produce enough insulin, then it is known as insulin deficiency. However, the body cannot use insulin, glucose stay in the blood and thus diabetes developed.

There are 3 types of diabetes namely Type1, Type 2 are commonly occurred and Gestational diabetes which is occurred during pregnancy time only.

Machine Learning Techniques like Supervised, Unsupervised, Semi-Supervised, Reinforcement, Evolutionary learning, and deep learning algorithms, etc. are useful for predicting diabetes in earlier stages.

2. Literature Review

A. Related Works

The related work done by the various researchers for predicting the health care applications [1]-[3], [6]-[9]. This research focuses particularly for the prediction of diabetic disease using different datasets has been analyzed. The results obtained have been analyzed.

Kayaer, Kamer, and Tulay Yldrm, [4] implemented a Medical diagnosis on Pima Indian diabetes using general regression neural networks.

Choubey, Dilip Kumar, et al., [5] developed a model for Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection.

Priyadarshini, R., Dash, N., & Mishra, R. [10] invented a Novel approach to predict diabetes mellitus using modified Extreme learning machine.

Beloufa, F., & Chikh, M. A. [11] designed a fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm.

Christopher, J. J., Nehemiah, H. K., & Kannan, A., [12] developed a diabetes prediction model using swarm optimization approach for clinical knowledge mining.

Lekkas, S., & Mikhailov, L., [13] have been evolved with fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases.

Rajiv Kannan et al [14]-[16] discussed about feature selection and prediction of disease using machine learning approaches for health care applications.

In current scenarios, different type of results has been shown for a patient due to the variations in medical diagnosis methods. It is desirable to classify a patient should be a diabetic or non-diabetic category. The errors in diagnosis may lead to unnecessary treatments or no treatments at all when required. To overcome these errors, it is necessary to develop a model using machine learning techniques which will provide more accurate results and save the patients.

3. Overview of the Prediction Model

The overview of the diabetic prediction model using machine learning algorithm is as given in the Figure 1.

Firstly, the data collected is to be applied for preprocessing techniques to remove noisy data. Then it is split into 80% and 20% for training and testing process respectively. Secondly the prediction model is to be developed by applying appropriate machine learning techniques. Finally, testing data is applied to

*Corresponding author: aparna.k.s1998@gmail.com the model to evaluate the build model. The error occurred and performance metrics have been analyzed for training and testing datasets.

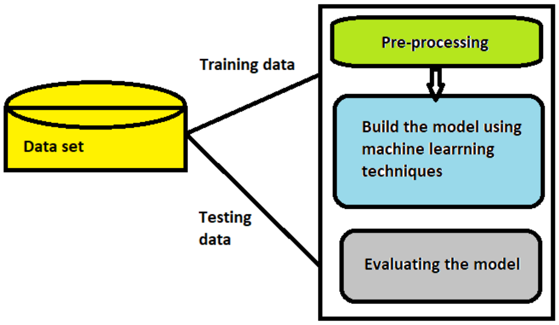

Fig. 1. Overview of the prediction model

4. Data Preprocessing

The data is to be preprocessed before applying machine learning algorithms for predicting diabetic patients.

5. Dataset Used

A. *PIMA Dataset*

The machine learning algorithms used for predicting the diabetic disease in earlier stage can be tested with the PIMA dataset available with open access [17]. It consists of 768 records of data with 268 diabetic patients and 500 non-diabetic patients from a population near Phoenix, Arizona, USA. It has the following attributes as listed below:

1. Pregnancy - Number of times pregnant
2. Glucose - glucose concentration at 2h
3. Blood Pressure - Diastolic blood pressure
4. Skin Thickness - Triceps skin thickness
5. Insulin - 2h serum insulin
6. Body Mass Index- BMI
7. Diabetic Prediction Function – DPF
8. Age - Age of the patient
9. Outcome - (1 for diabetic and 0 for non-diabetic)

B. *Vanderbilt Dataset*

The Vanderbilt dataset consists of 390 records containing 18 features for of 330 non-diabetic and 60 diabetic patients downloaded from data.world [18]. It is collected by surveying hundreds of rural African-American patients with different parameters. The dataset include the following features as listed below:

1. Cholestrol
2. Glucose
3. HDL Chol
4. Chol/HDL ratio
5. Age
6. Gender
7. Height
8. Weight
9. BMI
10. Systolic BP
11. Diastolic BP
12. Waist
13. Hip
14. Patient number
15. Waist/hip ratio and two unnamed features

C. *Preprocessing and Normalization*

In general, the features with zero values or unknown values may leads to reduce the performance of prediction model. Since these may be having primary factors to predict the disease. The PIMA dataset contains zero values for few of the attributes as listed below: Glucose (5), Blood Pressure (35), Skin thickness (227), Insulin (374) and BMI (11). These zeros or missing values can be filled in many ways. Here, such values in the records to be filled with the numeric mean value of the corresponding feature.

Similarly, for the Vanderbilt dataset, the missing values are deleted and duplicates are removed.

For normalizing both datasets, min-max normalization is applied to get all the values lies between 0 and 1. Hence the processing of large dataset can be done easily.

Feature Selection:

If all the attributes in the dataset is used for prediction model, it leads to erroneous results. To avoid that, the significant features among all features can be extracted using several algorithms. The following list shows few of the algorithms used for selecting the best features or the critical factors in the dataset.

1. Greedy Step wise Search Algorithm
2. ANOVA
3. Mutual Information
4. Genetic Algorithm
5. Pearsons Coefficient

6. Methods

Machine learning algorithms are classified into major three categories (i) Supervised (ii) Unsupervised (iii) Deep Learning methods as given in figure 2.

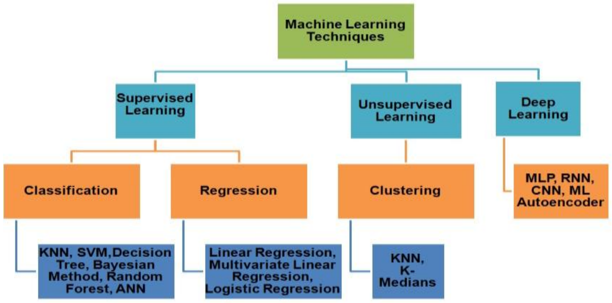
Machine Learning Techniques
Supervised Learning
Unsupervised Learning
Deep Learning
Classification
Regression
Clustering
MLP, RNN, CNN, ML Autoencoder
KNN, SVM,Decision Tree, Bayesian Method, Random Forest, ANN
Linear Regression, Multivariate Linear Regression, Logistic Regression
KNN, K-Medians

Fig. 2. Categories of Machine Learning algorithms

The algorithms used for predicting diabetics are listed below:

- SVM
- Logistic Regression
- Adaboost.M1
- Random Forest
- IBK
- J48graft
- Naïve Bayes

Table 1
Performance of various ML algorithms with feature selection methods

Data set used	Feature Selection Method	Features selected	Classification Algorithm used	Accu racy%	References
DS1	GSSA	2, 6, 8, 7	Multilayer perceptron	85.15	[1]
DS1	GSSA	1, 2, 6, 8, 7	Logistic regression	77.08	[1]
DS1	GSSA	1, 2, 6, 8, 7	Multilayer Perceptron	75.39	[1]
DS1	GSSA	1, 2, 6, 8, 7	Random Forest	75	[1]
DS1	PC	2, 6, 5, 1, 8	Decision tree	73.14	[1]
DS1	PC	2, 6, 5, 1, 8	Random Forest	77.14	[1]
DS1	PC	2, 6, 5, 1, 8	Naïve Bayes	78.28	[1]
DS1	PC	2, 6, 5, 1, 8	Linear Regression	78.85	[1]
DS1	PC	2, 6, 5, 1, 8	KNN	79.42	[1]
DS1	PC	2, 6, 5, 1, 8	AB	79.42	[1]
DS1	PC	2, 6, 5, 1, 8	SVM	77.71	[1]
DS2	ANOVA	2, 5	Logistic Regression	89.66	[3]
DS2	ANOVA	2, 5, 4 , 1, 12	Naïve Bayes	92.91%	[3]
DS2	ANOVA	2, 5, 4, 1, 12, 15 , 10, 8, 9	Stochastic Gradient Descent	90.68	[3]
DS2	ANOVA	2, 5, 4	KNN	92.65%	[3]
DS2	ANOVA	2, 5, 4, 1, 12	Decision Tree	88.54	[3]
DS2	ANOVA	2, 5, 4	Random Forest	92.65	[3]
DS2	ANOVA	2, 5	Support Vector Machine	92.56	[3]
DS2	MI	2, 13, 9	Logistic Regression	89.31	[3]
DS2	MI	2, 13, 9	Naïve Bayes	92.82	[3]
DS2	MI	2, 13, 9	Stochastic Gradient Descent	91.54	[3]
DS2	MI	2	KNN	92.30	[3]
DS2	MI	2	Decision Tree	88.12	[3]
DS2	MI	2, 13	Random Forest	92.22	[3]
DS2	MI	2	Support Vector Method	92.39	[3]
DS2	GA	1, 2, 3, 4, 5, 7, 9, 10, 11, 13	Genetic Algorithm	89.42	[3]
DS2	GA	1, 2, 11, 13	Naïve Bayes	93.59	[3]
DS2	GA	2, 5, 8, 9, 12	Stochastic Gradient Descent	93.58	[3]
DS2	GA	1, 2,7, 8,13	KNN	93.91	[3]
DS2	GA	1, 2, 11, 13	Decision tree	90.06	[3]
DS2	**GA**	**1,2,4, 10, 8, 13**	**Random Forest**	**93.95**	[3]
DS2	GA	2, 6	Support Vector Machine	93.27	[3]

- Stochastic Gradient Descent (SGD)
- KNN
- Multilayer Perceptron
- Decision Tree
- Linear Regression
- Genetic Algorithm
- GRNN
- ELM
- Artificial Bee Colony
- Swarm Intelligence
- Fuzzy Rule
- K- Means
- Neuro Fuzzy Inference
- Deep Learning Algorithm
- XGBoost (an Ensemble Model)

The survey has been done to know the performance of different machine learning algorithms. The performance of various machine learning algorithms alongwith various feature selection approaches applied on different datasets are tabulated. For example, accuracy obtained by the algorithms Logistic Regression, Naïve bayes, Stochastic Gradient Descent, KNN, Decision Tree, Random Forest and Support Vector Machine used alongwith ANOVA feature selection are tabulated as shown in Table 1. Also the features selected for that specific algorithm are shown in it. Other methods and the accuracy obtained for different datasets are also tabulated in it.

Where DS1 refers PIMA dataset & DS2 refers Vanderbilt dataset, GSSA- Greedy Stepwise Search Algorithm, PC – Pearson Coefficient, MI – Mutual Information, GA- Genetic Algorithm.

7. Results

It is seen that various features have been selected by using different feature selection approaches. It significantly impacts the performance of the model used. From Table 1, it is obvious that Random Forest with Genetic Algorithm outperforms with the accuracy 93.95 % when compared other algorithms.

8. Conclusion

The Earlier detection of diabetic disease can protect the patients from severe impacts of it. The PIMA dataset and Vanderbilt dataset are mainly considered for analyzing the performance of the diabetic prediction algorithms. The feature extraction methods are used to extract the best features from the datasets. After normalizing different machine learning algorithms have been applied on various datasets to know the performance of them. From the survey results, the Random forest algorithm with genetic algorithm proven the best among the other algorithms with the greatest accuracy.

References

[1] Sajratul Yakin Rubaiat, Md Monibor Rahman, Md. Kamrul Hasan, “Important Feature Selection and Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection”, (ICIET2018), 2018.

[2] Riihimaa P, "Impact of machine learning and feature selection on type 2 diabetes risk prediction", review Article, Journal of Medical Artificial Intelligence 2020, 3, 10.

[3] Simran Gill and Parthmesh Pathwar, "Prediction of Diabetes using various feature Selection and Machine Learning Paradigms", Easy Chair Preprint No. 6587, September 13, 2021.

[4] Kayaer, Kamer, and Tulay Yldrm. "Medical diagnosis on Pima Indian diabetes using general regression neural networks." (ICANN/ICONIP). 2003.

[5] Choubey, Dilip Kumar, et al. "Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection." (ICCCS 2016). 2017.

[6] Seera, M., & Lim, C. P., "A hybrid intelligent system for medical data classification. Expert Systems with Applications", 41(5), 2239-2249, 2014.

[7] Hayashi, Y., & Yukita, S., "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked", 2, 92-104, 2016.

[8] Kahramanli, H., & Allahverdi, N., "Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications", 35(1-2), 82-89, 2008

[9] Ahmad, A., Mustapha, A., Zahadi, E. D., Masah, N., & Yahaya, N. Y. "Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus", In Digital Information Processing and Communications, pp. 537-545, Springer, Berlin, Heidelberg, 2011.

[10] Priyadarshini, R., Dash, N., & Mishra, R., "A Novel approach to predict diabetes mellitus using modified Extreme learning machine", in Electronics and Communication Systems (ICECS), 2014 International Conference on, pp. 1-5, IEEE, 2014.

[11] Beloufa, F., & Chikh, M. A., "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm", Computer methods and programs in biomedicine, 112(1), 92-103, 2013.

[12] Christopher, J. J., Nehemiah, H. K., & Kannan, A., "A swarm optimization approach for clinical knowledge mining", Computer methods and programs in biomedicine, 121(3), 137-148, 2015.

[13] Lekkas, S., & Mikhailov, L., "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. Artificial Intelligence in Medicine, 50(2), 117-126, 2010.

[14] A. Rajivkannan, M. M. Kiruthiga, C. Praveen Kumar, A. Dhenaskumar, S. Manoj "Prediction of Parkinson's Disease using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT), Special Issue 2019.

[15] P. Renukadevi, A. Rajiv Kannan, "Improved Linear factor based Grasshopper Optimization algorithm with Ensemble Learning for Covid – 19 Forecasting", Neuroquantology, vol. 19, no. 8, pp. 169 – 181, August 2021.

[16] Vijayanand Sellamuthu Palanisamy, Rajiv Kannan Athiappan, Thirugnanasambandan Nagalingam, "Pap Smear based Cervical Cancer Detection using Residual Neural Networks Deep Learning Architecture", Wiley Online Library, 18 September 2021.

[17] PIMA data set: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[18] Vanderbilt dataset: http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets