Conference Proceeding Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Diabetes Disease Prediction Using Machine Learning On Big Data Of Healthcare:Survey Paper

[1] Sangeeta Waren,
[1]Research Scholar, Department of Computer Science & Engineering
Global Nature Care Sangathan's Group of Institutions, Jabalpur, M.P. , India

***Abstract*— Type 2 diabetes is one of the most common medical conditions in the world. several years in the last few years in the diagnosis of diabetes. the rise in difficulty of Deep Learning (DL) has encouraged researchers to try their hand at solving the tough problems So far, the model has achieved a 0.011 percent accuracy. The solution found employs a set of ML algorithms, however, many of which are rarely used in the context of this problem, so it is interesting to study their abilities with respect to diabetes prediction. In addition, there is no recent study that compares and reviews all of the combined modelling and methodology proposals. 6 years: This article covered all the ML and DL prediction strategies out there. Additionally, the application of rare ML classifiers to the Pima Indians had the aim of increasing their efficiency. Although the classifiers obtained a score of 68%-74%, This text proposes to use these classifiers in the field of diabetes prediction and incorporate them into a more comprehensive model.**

Keywords— Diabetes Mellitus, Big Data Analytics, Healthcare Machine Learning.

Introduction

Diabetes is one of the frequent diseases that targets the elderly population worldwide. According to the International Diabetes Federation, 451 million people across the world were diabetic in 2017. The expectations are that this number will increase to affect 693 million people in the coming 26 years [1]. Diabetes is considered as a chronic disease associated with an abnormal state of the human body where the level of blood glucose is inconsistent due to some pancreas dysfunction that leads to the production of little or no insulin at all, causing diabetes of type 1 or cells to become resistant to insulin, causing diabetes of type 2 [2,3]. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes. Even though it's incurable, it can be managed by treatment and medication. Individuals with diabetes face a risk of developing some secondary health issues such as heart diseases and nerve damage. Thus, early detection and treatment of diabetes can prevent complications and assist in reducing the risk of severe health problems. Many researchers in the bioinformatics field have attempted to address this disease and tried to create systems and tools that will help in diabetes prediction. They either built prediction models using different types of machine learning algorithms such as classification or association algorithms. Decision Trees, Support Vector Machine (SVM), and Linear Regression were the most common algorithms [4–6].

Artificial Neural Network (ANN) is another type of machine learning technique. It is well-known for its high performance and accuracy. Furthermore, due to the increasing size and complexity of the data, Deep Learning (DL) has been introduced as an improvement to ANN. Recent studies that have used DL produced remarkable results [7,8].

The accuracy rate produced by these methods varied. This has encouraged researchers to attempt improving the accuracy by either building models with classifiers that haven't been used or combine different classifiers [9–11]. The majority of the studies in the field of the Diabetes prediction used the public Pima Indian Dataset obtained from the UCI repository.

Some surveys have been published, but they are different from this study. For example, [12] discussed the well-known ML and DL techniques used in predicting diabetes. The authors described only the studies related to Decision Tree, Support Vector Machine, Artificial Neural Network, and some DL techniques. Reference [13] also surveyed the main well-known ML classification techniques to predict diabetes. Reference [14] investigated the ML techniques used in predicting different diseases, including diabetes, which is discussed by only five related works. Reference [15] also discussed the ML techniques used in predicting heart, breast cancer and diabetes diseases, where the authors discussed only two studies about diabetes prediction. In fact, all these surveys investigated the ML techniques used in predicting diabetes disease. However, only one survey [12] addressed the DL state-of-the art techniques that predicted diabetes disease. In addition, two reviews [12,13] focused only on the diabetes disease, unlike [14,15] which discussed several diseases including diabetes introduced by five studies in [14] and only two studies in [15].

This research paper discusses Machine and Deep Learning techniques in addition to combined models for the prediction of diabetes [16] published since 2013. Combined models are defined as a combination of two or more classifiers. For example, they can be a combination of two or more ML techniques, or ML with AI techniques. This paper provides a systematic review of the performances of different Machine and Deep Learning classifiers obtained from different papers collected in the last six years. To save space, only a sample of the studies have been reviewed in the related works section while Tables A1–A3 summarize all the existing studies in the last six years (more than 40 studies in total, see the Appendix A). Moreover, the frequencies of using the ML classifiers are calculated. Upon that, the popular classifiers are identified, and then the rarely (or not) used classifiers are applied on the PID using the Weka tool. Up to our knowledge, none of the existing surveys discussed these classifiers. Based on the results we acquired, a comparative analysis is performed with other research studies that used the same dataset. Finally, this paper aims to create a unified repository the researchers can refer to when they want to predict diabetes.

This paper is organized as follows: Section 2 describes the related works that used ML and DL techniques in addition to combined models. Section 3 addresses a comprehensive discussion about the related studies showing the main diabetes datasets and their features (summarized in Table A4) in addition to an overview of several ML/DL algorithms along with their advantages and disadvantages (summarized in Table A5). Section 4 presents one case study to predict the diabetes in addition to the results and the discussion about the performance of the classifiers used. Finally, Section 5 states the final findings and the conclusions of the study.

I. Literature Review

To perform this study, 27 Machine Learning related studies have been collected, as seen in Table A1 (see the Appendix A) [4,5,17–41]. However to save space, only 10 of the most recent studies have been discussed in details. In addition, seven studies related to Deep Learning techniques were found (see Table A2 in the Appendix A) and discussed in this section. Moreover, six papers presenting combined models were collected and presented in Table A3 (see the Appendix A) but they are not discussed. Tables A1 and A2 provide the reference, the year of publication, the evaluation measure with its obtained value, and the dataset used for each study published in the last six years. Moreover, the main datasets used in the below discussed studies are summarized in Table A4 (see the Appendix A) indicating their size and the main features utilized.

2.1. Related Works Using Machine Learning

ML algorithms are very well-known in the medical field for predicting diseases. Many researchers have used ML techniques to predict diabetes in an effort to obtain the best and most accurate results [16].

Kandhasamy and Balamurali [4] used multiple classifiers SVM, J48, K-Nearest Neighbors (KNN), and Random Forest. The classification was performed on a dataset taken from the UCI repository (for more details see Table A4). The results of the classifiers were compared based on the values of the accuracy, sensitivity, and specificity. The classification was done in two cases, when the dataset is pre-processed and without preprocessing by using 5-fold cross validation. The authors didn't explain the pre-processing step applied on the dataset, they just mentioned that the noise was removed from the data. They reported that the decision tree J48 classifier has the highest accuracy rate being 73.82% without pre-processing, while the classifiers KNN (k = 1) and Random Forest showed the highest accuracy rate of 100% after pre-processing the data.

Moreover, Yuvaraj and Sripreethaa [17] presented an application for diabetes prediction using three different ML algorithms including Random Forest, Decision Tree, and the Naïve Bayes. The Pima Indian Diabetes dataset (PID) was used after pre-processing it. The authors didn't mention how the data was pre-processed, however they discussed the Information Gain method used for feature selection to extract the relevant features. They used only eight main attributes among 13 (see Table A4). In addition, they divided the dataset into 70% for training and 30% for testing. The results showed that the random forest algorithm had the highest accuracy rate of 94%.

Furthermore, Tafa et al. [18] proposed a new integrated improved model of SVM and Naïve Bayes for predicting the diabetes. The model was evaluated using a dataset collected from three different locations in Kosovo. The dataset contains eight attributes and 402 patients where 80 patients had type 2 diabetes. Some attributes utilized in this study (see Table A4) have not been investigated before, including the regular diet, physical activity, and family history of diabetes. The authors didn't mention whether the data was pre-processed or not. For the validation test, they split the dataset into 50% for each of the training and testing sets. The proposed combined algorithms have improved the accuracy of the prediction to reach 97.6%. This value was compared with the performance of SVM and Naïve Bayes achieving 95.52% and 94.52%, respectively.

In addition, Deepti and Dilip [19] used Decision Tree, SVM, and Naive Bayes classifiers to detect diabetes. The aim was to identify the classifier with the highest accuracy. The Pima Indian dataset was used for this study. The partition of the dataset is done by means of 10-folds cross-validation. The authors didn't discuss the data preprocessing. The performance was evaluated using the measures of the accuracy, the precision, recall, and the F-measure. The highest

accuracy was obtained by the Naive Bayes, which reached 76.30%.

Mercaldo et al. [20] used six different classifiers. The classifiers are J48, Multilayer Perceptron, HoeffdingTree, JRip, BayesNet, and RandomForest. The Pima Indian dataset was also utilized for this study. The authors didn't mention a preprocessing step, however, they employed two algorithms, GreedyStepwise and BestFirst, to determine the discriminatory attributes that help in increase the classification performance. Four attributes have been selected, namely body mass index, plasma glucose concentration, diabetes pedigree function, and age. A 10 fold-cross validation is applied to the dataset. The comparison between the classifiers was made based on the value of the precision, the recall, and the F-Measure. The result showed the precision value equals to 0.757, recall equals to 0.762, and F-measure equals to 0.759 using the Hoeffding Tree algorithm. This is the highest performance compared to the others.

In addition to the other studies, Negi and Jaiswal [21] aimed to apply the SVM to predict diabetes. The Pima Indians and Diabetes 130-US datasets were used as a combined dataset. The motivation of this study was to validate the reliability of the results as other researchers often used a single dataset.

The dataset contains 102,538 samples and 49 attributes where 64,419 were positive samples and 38,115 were negative samples. The authors didn't discuss the attributes used in this study. The dataset is pre-processed by replacing the missing values and out of range data by zero, the non-numerical values are changed to numerical values, and finally the data is normalized between 0 and 1. Different feature selection methods were used prior to the application of the SVM model. The Fselect script from LIBSVM package selected four attributes, while Wrapper and Ranker methods (from Weka Tool) selected nine and 20 attributes, respectively. For the validation process, the authors used 10-fold cross validation technique. By using a combined dataset, the diabetes prediction might be more reliable, with an accuracy of 72%.

Moreover, Olaniyi and Adnan [22] used a Multilayer Feed-Forward Neural Network. The back-propagation algorithm was used for training the algorithm. The aim was to improve the accuracy of diabetes prediction. The Pima Indian Diabetes database was used (see Table A4). The authors normalized the dataset before processing to the classification in order to obtain a numerical stability. It consisted of dividing each sample attributes by their corresponding amplitude to make all the dataset values between 0 and 1. After that, the dataset is divided into 500 samples for a training set and 268 for the testing set. The accuracy obtained was 82% which is considered as a high accuracy rate.

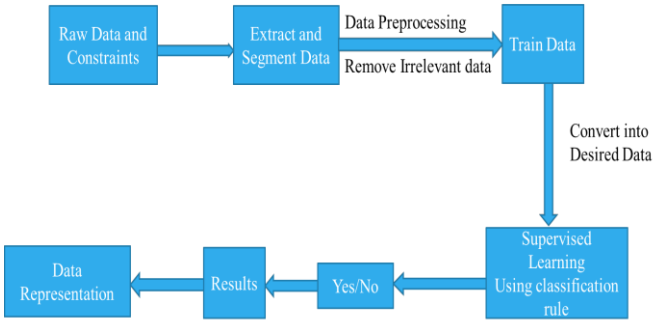
Soltani and Jafarian [23] used the Probabilistic Neural Network (PNN) to predict diabetes. The algorithm was applied to the Pima Indian dataset. The authors didn't apply any pre-processing technique. The dataset is divided into 90% for the training set and 10% for the testing set. The proposed technique achieved accuracies of 89.56%, 81.49% for the training and testing data, respectively.

Rakshit et al. [24] used a Two-Class Neural Network to predict diabetes using the Pima Indian dataset. The authors pre-processed the dataset by normalizing all the sample attributes values using the mean and the standard deviation of each attribute in order to obtain a numerical stability. In addition, they extracted the relevant features using the correlation. However, the authors didn't mention these discriminatory features. The dataset was split into a training set containing 314 samples and a testing set comprising 78 samples. The result of this model achieved the highest accuracy of 83.3% when compared to other accuracies obtained from the previous studies.

Mamuda and Sathasivam [25] applied three supervised learning algorithms including Levenberg Marquardt (LM), Bayesian Regulation (BR), Scaled Conjugate Gradient (SCG). This study used the Pima Indian dataset (with 768 samples and eight attributes, see Table A4) for evaluating the performance. For the validation study, the 10-fold cross validation was used to split the data into training and testing. The authors reported that Levenberg Marquardt (LM) obtained the best performance on the validation set based on the Mean Squared Errorr (MSE) equals to 0.00025091.

II. Proposed Work

The Proposed method use KNN algorithm for classification and prediction of diabetes using trained data. And, the proposed system also predicts the time of getting diabetes.



Data Constraints

Data is a collection global dataset. IN this system use Pima Indian data set is used for training a model. Data set contain

21 parameters and around 1000 dataset. The dataset feature/parameters are:

- Age
- Gender
- Relation
- DOB
- Sugar tested value
- Symptoms
- Family history etc.

This are data is trained to the model for the prediction of diabetes.

Train Dataset and Test Dataset

The training data is a initial set of data which is used to understand the program. This is the one in which we have to train the model first because to set the feature and this data is available on system. This data is used to teach the machine for do different actions. It is the data in which model can learn with algorithm to teach the model and doing work automatic.

Testing data is the input given to a software. It shows the data affects when the execution of the module that specifying and this is basically used for testing.

Pre-processing of data

Data preprocessing is a process in which that is actual use for converting the basic data into the clean data set. It is the step in which the data transform or an encode to the state that the machine can be easily parse. The major task of data preprocessing in learning process is to remove the unwanted data and filling the missed value. So that it help to machine can be trained easily.

Figure 2: Data Pre-processing.

Feature Extraction

Feature Extraction is the method in which it used for alter the key data for features of outcomes. This, trait square is used to compute the characteristics of designs given that facilitate in different amid the class of key pattern details. This method involving to decrease the counts of resource required to describe the huge set of data. Feature extraction is an attribute reduction process. This is also used to increasing the speed and effectiveness of supervised learning.

ML Algorithm: KNN

The k-nearest neighbor's is a ML algorithm is the non-parametric method proposed by Thomas Cover used for Regression and Classification. This algorithm is mainly used for the classification of problems in the industry. KNN algorithm is a type of instance-based learning method. This algorithm relies on the distance for objects classification, training data normalizing to the improve its accuracy dramatically. The neighbors are derived from the set of things for which classes or object property values are known. It can be thought of as a training set for the algorithm, although no explicit training steps are required.

System Design

Designing of system is the process in which it is used to define the interface, modules and data for a system to specified the demand to satisfy. System design is seen as the application of the system theory. The main thing of the design a system is to develop the system architecture by giving the data and information that is necessary for the implementation of a system. In this project three-tier architecture is used.

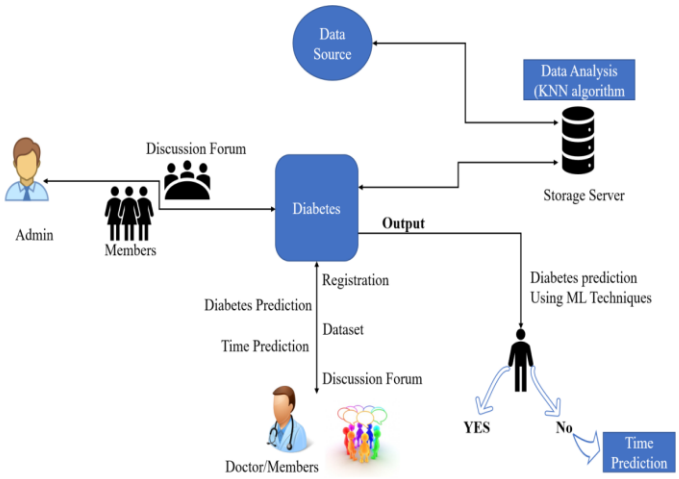


Figure : Architecture Design.

Moreover, one of the classifiers under the function's category has got a significant accuracy of 72.14%. It was attained by the SMO classifier. Sequential Minimal Optimization (SMO) is used in training the Support Vector Machines (SVM). This is due to the need for a solution to the quadratic programming (QP) optimization problem in order to train the SVM. SMO breaks the QP into small and manageable problems that can be solved in less time. Also, the amount of memory required to handle the small problems is minimized. This allows SMO to process very large datasets. Moreover, it has pre-processing abilities in which it replaces the missing values as well as transforming nominal values into binary ones. Furthermore, it performs normalization on the data by default. This helps in boosting the prediction accuracy [63].

For the Bayes category, the Bayes Nets gave remarkable results in many fields such as aircraft systems, scientific researches, and public safety [64]. Although, it performs well, it is not recommended to be used in prediction problems such as this study. This is due to the fact that this algorithm looks for the impact of the variables on the result. It outperforms regression functions when it comes to determine the effect of the variables.

The least accuracy was produced by the KStar classifier. It can handle noisy data and it requires less time to train the data. However, its performance becomes better with large datasets. Also, to use this technique, the value of the parameter k needs to be defined. The computation cost is very high as it needs to calculate the distance of the instances in the training sample [65].

As a summary, the decision tree algorithms obtained the highest accuracy and it is recommended to be used in the classification and prediction problems. The other algorithms have also a competitive accuracy. Hence, we recommend using these algorithms in the classification and prediction studies to take benefit from their strengths. Moreover, these algorithms can be used in a combined model with other Deep or Machine Learning techniques as well as Artificial Intelligence techniques to boost their accuracy.

III. Results & Analysis

Researchers are passionate to try different types of classifiers and build new models with an effort to enhance the accuracy of diabetes prediction. In this paper, the same vision was

followed to reach high prediction accuracy. All the Machine Learning (ML) and Deep Learning (DL) classifiers that have been used in the last six years were reviewed regarding their frequency of use and accuracy. ML classifiers with one or zero frequency have been implemented on the PID dataset to set recommendations regarding their usage. The obtained accuracy by these ML techniques was 68%–74%. For the DL algorithms, the highest accuracy achieved by researchers was 95%. As a future work, the non-used classifiers can be applied to other datasets in a combined model to enhance further the accuracy of predicting the Diabetes disease.

Implementation can be described as the realization of an application, or execution of the plans, ideas, models, design and system development, specification of the model, standard, algorithms used in the system, or authority. In computer science, an implement is explained as the realization of technically specified or algorithms' as a programed, a software component, or any others computer systems through computer programming and deployment. Many of the implementations may existed for a given specification or standard.

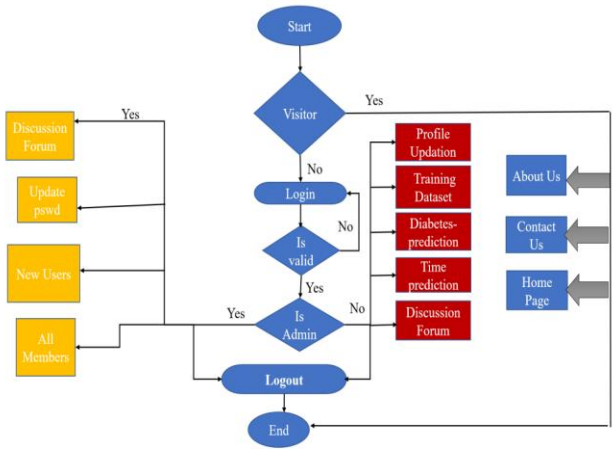


Figure: Control Flow

IV. Conclusion & Future Work

The prediction of diabetes is one the of great importance in today scenario, and concerning with its severe complications. Due to the biggest reason for the death in worldwide is diabetes. The System model is mainly focus to identification of diabetes using some of the parameters. System is useful to physicians to predict the diabetes in initial dais. So, that conventional treatments and solutions may be given to the patients. System used some of the techniques like ML for the prediction, so that to get the more precise results. There have been fortune of investigation on the diabetes imprint. Building diabetes disease prediction system is useful for hospitals and doctors. System predicts disease at early stages, so doctors can treat patients in a better way. Proposed model is the real time application in which is meant for multiple hospitals and predicts disease in less time. As we use machine learning algorithms for disease prediction, we will get more accurate and efficient results.

V. References

1. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; Fernandes, J.D.R.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pr. 2018, 138, 271–281. [CrossRef] [PubMed]
2. Sanz, J.A.; Galar, M.; Jurio, A.; Brugos, A.; Pagola, M.; Bustince, H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Appl. Soft Comput. 2014, 20, 103–111. [CrossRef]
3. Varma, K.V.; Rao, A.A.; Lakshmi, T.S.M.; Rao, P.N. A computational intelligence approach for a better diagnosis of diabetic patients. Comput. Electr. Eng. 2014, 40, 1758–1765. [CrossRef]
4. Kandhasamy, J.P.; Balamurali, S. Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Comput. Sci. 2015, 47, 45–51. [CrossRef]
5. Iyer, A.; Jeyalatha, S.; Sumbaly, R. Diagnosis of Diabetes Using Classification Mining Techniques. Int. J. Data Min. Knowl. Manag. Process. 2015, 5, 1–14. [CrossRef]
6. Razavian, N.; Blecker, S.; Schmidt, A.M.; Smith-McLallen, A.; Nigam, S.; Sontag, D. Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors. Big Data 2015, 3, 277–287. [CrossRef]
7. Ashiquzzaman, A.; Kawsar Tushar, A.; Rashedul Islam, M.D.; Shon, D.; Kichang, L.M.; Jeong-Ho, P.; Dong-Sun, L.; Jongmyon, K. Reduction of overfitting in diabetes prediction using deep learning neural network. In IT Convergence and Security; Lecture Notes in Electrical Engineering; Springer: Singapore, 2017; Volume 449.
8. Swapna, G.; Soman, K.P.; Vinayakumar, R. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Comput. Sci. 2018, 132, 1253–1262.
9. Rahimloo, P.; Jafarian, A. Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them. Bull. Société R. Sci. Liège 2016, 85, 1148–1164.
10. Gill, N.S.; Mittal, P. A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. J. Theor. Appl. Inf. Technol. 2016, 87, 1–10.
11. NirmalaDevi, M.; Alias Balamurugan, S.A.; Swathi, U.V. An amalgam KNN to predict diabetes mellitus. In Proceedings of the 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, India, 25–26 March 2013; pp. 691–695.
12. Sun, Y.L.; Zhang, D.L. Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey. Teh. Vjesn. 2019, 26, 872–880.
13. Choudhury, A.; Gupta, D. A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. In Recent Developments in Machine Learning and Data Analytics; Springer: Singapore, 2019; pp. 67–68.
14. Meherwar, F.; Maruf, P. Survey of Machine Learning Algorithms for Disease Diagnostic. J. Intell. Learn. Syst. Appl. 2017, 9, 1–16.
15. Vijiyarani, S.; Sudha, S. Disease Prediction in Data Mining Technique—A Survey. Int. J. Comput. Appl. Inf. Technol. 2013, 2, 17–21.
16. Deo, R.C. Machine Learning in Medicine. Circulation 2015, 132, 1920–1930. [CrossRef] [PubMed]
17. Yuvaraj, N.; SriPreethaa, K.R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust. Comput. 2017, 22, 1–9. [CrossRef]
18. Tafa, Z.; Pervetica, N.; Karahoda, B. An intelligent system for diabetes prediction. In Proceedings of the 2015

4th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 14–18 June 2015; pp. 378–382.

19. Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 2018, 132, 1578–1585. [CrossRef]

20. Mercaldo, F.; Nardone, V.; Santone, A. Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Comput. Sci. 2017, 112, 2519–2528. [CrossRef]

21. Negi, A.; Jaiswal, V. A first attempt to develop a diabetes prediction method based on different global datasets. In Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 22–24 December 2016; pp. 237–241.

22. Olaniyi, E.O.; Adnan, K. Onset diabetes diagnosis using artificial neural network. Int. J. Sci. Eng. Res. 2014, 5, 754–759.