Ann. For. Res. 65(1): 7529-7539, 2022
ISSN: 18448135, 20652445

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

A REVIEW ON DIABETES PREDICTION USING MACHINE LEARNING ON HEALTHCARE BIG DATA

Dr. Ashish Mishra
Professor, Dept of CSE Gyan Ganga Institute of Technology and Sciences, Jabalpur (M.P.)
ashishmishra@ggits.org

Isha Choubey
Research Scholar, Dept of CSE Gyan Ganga Institute of Technology and Sciences, Jabalpur (M.P.) ishachoubey28@gmail.com

Abstract— Diabetes of type 2 is one of the world's most frequent medical disorders. Diabetes diagnosis for numerous years in recent years. Deep Learning's (DL) increase in difficulty has prompted researchers to try and solve the difficult challenges To now, the model has attained a precision of 0.011 percent. The approach obtained uses a set of ML algorithms, many never utilised to address this problem. It is therefore fascinating to explore their ability to forecast diabetes. Furthermore, no current work compares and reviews all combined proposals for modelling and technique. 6 years: 6 years: This article discussed all of the strategies for ML and DL prediction. In addition, the purpose of applying unusual ML classifications to the Pima Indians was to improve their efficiency. While the classifiers have attained a rating of 68-74 percent, this paper recommends the use of these classifiers in a more complete diabetes prediction.
Keywords— Diabetes Mellitus, Big Data Analytics, Healthcare Machine Learning.

I. INTRODUCTION

One of the many illnesses that frequently affect the elderly globally is diabetes. According to the International Diabetes Federation, there were 451 million diabetics worldwide in 2017. Within the next 26 years,it is guessed that this number would ascend to 693 million [1]. Type 1 diabetes is considered a persistent illness brought about by a strange condition of the human body where the blood glucose level differs due to pancreatic brokenness that results in the formation of close to zero insulin or cells that become impervious to insulin, which causes type 2 diabetes[2,3]. Scientists believe that environmental and lifestyle variables have a big impact on the condition, even if the actual cause of diabetes is still unknown. Despite being incurable, it is frequently treated and treated with medicine. Diabetes patients run the danger of experiencing additional health problems like heart disease and nerve damage. Therefore, distinguishing and treating diabetes early can assist with staying away from difficulties and lower the gamble of serious medical problems. The disease of diabetes has been the focus of numerous bioinformatics researchers who have tried to develop tools and systems to aid in diabetes prediction. Utilizing different AI techniques, for example, arrangement or affiliation calculations, they either made expectation models. The most well known strategies were rectilinear regression, decision trees, and support vector machines (SVM) [4-6].

One more kind of AI strategy is the artificial neural network (ANN). It is renowned for being highly effective and precise.Moreover, Due to the rising size and intricacy of information, deep learning (DL) has been introduced as an improvement over ANN. Remarkable outcomes were obtained in recent investigations that used DL [7,8].
These techniques achieved varying accuracy rates.This has provoked scholastics to chip away at expanding exactness by either making models with classifiers that haven't been utilized or by joining different classifiers [9-11]. The public Pima Indian Dataset recovered from the UCI vault was utilized in most concentrations in the field of diabetes expectation.

One of the many illnesses that frequently affect the elderly globally is diabetes. According to the International Diabetes Federation, there were 451 million diabetics worldwide in 2017. Within the next 26 years, it is anticipated that this number would rise to 693 million [1]. Type 1 diabetes is considered a constant illness brought about by an unusual condition of the human body where the blood glucose level varies because of pancreatic brokenness that outcomes in the development of practically no insulin or cells that become safe to insulin, which causes type 2 diabetes[2,3]. Albeit the specific etiology of diabetes is at this point unclear, researchers believe that both innate and ecological way of life factors fundamentally affect the infection. Despite being incurable, it is frequently treated and treated with medicine. Diabetes patients run the danger of experiencing additional health problems like heart disease and nerve damage. Therefore, detecting and treating diabetes early can help stay away from inconveniences and lower the gamble of serious medical problems. The sickness of diabetes has been the focal point of various bioinformatics scientists who have attempted to foster apparatuses and frameworks to support diabetes forecasts. They either made gauge models using different artificial intelligence strategies, similar to course of action or connection estimations. Rectilinear backslide, decision trees, and sponsorship vector machines were the most often used techniques.SVM) [4-6].
One more sort of machine learning method is the artificial neural network (ANN). It is renowned for being highly effective and precise. Furthermore, Deep Learning (DL) Deep Learning (DL) has been introduced as an advancement over ANN due to the growing size and complexity of the data. Remarkable outcomes were obtained in recent investigations that used DL [7,8].
These techniques achieved varying accuracy rates. This has motivated scholars to work on improving precision by either creating models with underutilized classifiers or by merging several classifiers[9–11]. The public Pima Indian Dataset recovered from the UCI store was utilized in most concentrations in the field of diabetes expectation.
There have been published surveys, but they are not the same as this one. For instance, [12] covered the notable ML and DL strategies applied in diabetes expectation. Only experiments including Decision Tree, Support Vector Machine, Artificial Neural Network, and a few DL approaches were described by the authors. Reference [13] also reviewed the most popular machine learning (ML) diabetes prediction algorithms. Diabetes, which is covered by only five linked studies, was one of the diseases covered by reference [14], which looked into the ML approaches used in

forecasting various diseases. Reference [15],while the creators just covered two examinations about diabetes forecast, likewise analyzed the ML approaches utilized in anticipating heart, disease, and diabetes issues. In fact, one of these studies looked into how machine learning techniques were utilised to predict diabeticinfection. In any case, just a single report [12] examined the best in class DL methods for predicting the development of diabetic illness. Additionally, compared to [14,15] which examined a variety of disorders remembering diabetes presented by five investigations for
[14] and only two examinations in [15,21], two surveys [12,13] zeroed in just on the diabetes illness.

This exploration study talks about machine and profound learning strategies as well as incorporated models for the expectation of diabetes [16] that were released starting in 2013. A minimum of two classifiers are combined to form consolidated models. For instance, they will combine at least two ML techniques or ML with artificial intelligence methodologies. This document delivers a logical overview of the demonstrations of several Machine and Profound Gaining classifiers culled from many studies over the previous six years. The frequency of applying the ML classifiers is also established. Following that, Identification of the learning classifiers and the infrequently (or never) used classifiers are then applied to the PID using the Weka apparatus. As far as anyone knows, none of the major reviews have looked at these classifiers. In support of the findings we found, a comparable inquiry was conducted using additional research focuses that made use of the same dataset. Finally, the goal of this research is to compile a resource that experts can refer to when predicting diabetes.

The course of action of this paper is as per the following: Region 2 shows the connected exploration that integrates both connected models and recently utilized ML and DL strategies. Notwithstanding a description of a few ML/DL calculations and their advantages and disadvantages, Segment 3 tends to be a lengthy discussion regarding the connected studies demonstrating the majority of diabetes datasets and their components (summed up in Table A5). Segment 4 includes a contextual inquiry to forecast diabetes, along with the findings and a discussion of the classifiers' display and use. Finally, Area 5 communicates a conclusive discovery and the review's conclusion.

II. LITERATURE REVIEW

Table A1 (see Appendix A) lists 27 studies on machine learning that were gathered for this investigation [4,5,17–41]. However, just 10 of the most current research have been thoroughly discussed in order to conserve space. Also discovered and discussed in this section are seven examinations relating to Deep Learning methods(see Table A2 in Appendix A). Additionally, Table A3 (see Appendix A) in the collection of six publications giving combined models contains these papers, but they are not mentioned.Each study distributed over the most recent six years incorporates the reference, the time of distribution, the appraisal measure and its gotten esteem, and the dataset employed in Tables A1 and A2. Additionally, Table A4 (see the Appendix A)

provides a summary of the primary datasets utilized in the examinations introduced beneath, including their size and key properties.

2.1. Additional Machine Learning-Related Works

In the world of medicine, ML algorithms are well-known for their ability to predict disease.
To achieve the best and most precise results, many specialists have utilized ML procedures to anticipate diabetes[16].
Multiple classifiers, including SVM, J48, K-Nearest Neighbors (KNN), and Random Forest, were utilised by Kandhasamy and Balamurali [4]. A dataset from the UCI repository was used for the categorization (for more details see Table A4). In light of the benefits of the exactness, awareness, and particularity, the effects of the classifiers were examined. When the dataset was pre-handled and without preprocessing, the characterisation was completed in two cases using a 5- overlap cross-approval. The designers failed to explain the pre-handling procedure used on the dataset; nonetheless, they have recently stated that the commotion was removed from the data. According to their findings,the KNN (k = 1) and Random Forest classifiers had the best exactness paces of 100 percent after pre-handling the information, while the choice tree J48 classifier had the most noteworthy precision pace of 73.82% without it.
Additionally, Yuvaraj and Sripreethaa [17] proposed a diabetes expectation application utilizing three unique ML calculations, including Random Forest, Decision Tree, and Nave Bayes. After pre-handling, the Pima Indian Diabetes dataset (PID) was used The authors mentioned the Data Gain approach for highlight choice used to remove the relevant attributes, but they made no notice of how the information was pre-handled. Out of the 13 available qualities, only eight were used (see Table A4). They also split the dataset into 30% for testing and 70% for training. With a 94% accuracy rate, the random forest algorithm provided the best results.
A new incorporated better model joining SVM and Nave Bayes for anticipating diabetes was also proposed by Tafa et al. [18]. A dataset gathered from three separate Kosovo areas was utilized to assess the model. Eight credits and 402 patients are considered in the sample, of whom 80 had type 2 diabetes. A few of the credits used in this evaluation (see
Table A4), such as the typical eating pattern, active work, and family history of diabetes, have not previously been looked at. The pre-handling of the information was not referenced by the creators. For the preparation and testing sets for the approval test, they separated the dataset into equal parts. The proposed joined techniques have expanded forecast exactness to 97.6%. This value was compared to SVM and Naive Bayes performance, which achieved 95.52% and 94.52%, respectively.
Furthermore, Deepti and Dilip [19] detected diabetes using Decision Tree, SVM, and Naive Bayes classifiers. Finding the classifier with the maximum accuracy was the goal.For this examination, the Pima Indian dataset was used. Ten times cross-approval is utilized to parcel the dataset. The researchers didn't want to talk about how the data were prepared. Performance was measured using the F-measure, recall, accuracy, and precision. The Naive Bayes model achieved the highest level of precision (76.30%).

Six distinct classifiers were utilised by Mercaldo et al. [20]. J48, Multi-facet Perceptron, Hoeffding Tree, JRip, BayesNet, and RandomForest are the classifiers. For this review, the Pima Indian dataset was also used. Although the creators didn't indicate a pretreatment step, they utilized the equations GreedyStepwise and BestFirst to choose the one-sided credits that assist to further develop the gathering execution. The four ascribes—specifically, weight file, plasma glucose focus, diabetic family capability, and age—have been chosen. The dataset is cross-approved with a 10 overlay. The accuracy, review, and F-Measure values were taken into consideration while comparing the classifiers. The Hoeffding Tree algorithm produced a precision value of 0.757, recall value of 0.762, and F-measure value of 0.759. Contrasted with the others, this show is the most grounded.

Negi and Jaiswal [22] tried to apply the SVM to foresee diabetes notwithstanding the other examination. The Diabetes 130-US and Pima Indians datasets were integrated to create a single dataset. Since other researchers frequently used the same dataset, the study's objective was to confirm that the findings were of consistently high quality.

The dataset includes 49 characteristics and 102,538 samples, of which 64,419 were positive examples and 38,115 were negative examples . The attributes utilized in this study were not talked about by the creators. The dataset is pre- processed by converting the non-numerical values to numerical values, supplanting missing qualities and out-of-range information with zero, and normalizing the data between 0 and 1. Before the SVM model was applied, various feature selection techniques were employed. While the Wrapper and Ranker methods (from the Weka Tool) picked nine and twenty qualities, separately, the Fselect script from the LIBSVM bundle just chose four credits. The creators utilized a 10-overlay cross approval procedure for the approval cycle. The diabetes expectation may be stronger by using a combined dataset, with an accuracy of 72%.

Additionally, Olaniyi and Adnan's [23] Multilayer Feed- Forward Neural Network was employed. The calculation was prepared utilizing the back-engendering approach. The objective was to

raise the precision of diabetes forecasts. It made use of the Pima Indian Diabetes database. To ensure mathematical soundness, the creators standardized the dataset before handling the characterization. It involved dividing up each example's attributes according to how well they compare, giving every dataset a value between 0 and 1. The dataset is then parted into 268 examples for the testing set and 500 examples for the preparation set. The precision pace of 82% that was accomplished is viewed as high.

III. GENERALIZED METHODOLOGY

The Proposed strategy uses KNN calculation for grouping and expectation of diabetes utilizing prepared information. Additionally, the suggested system analyzes when someone might get diabetes..

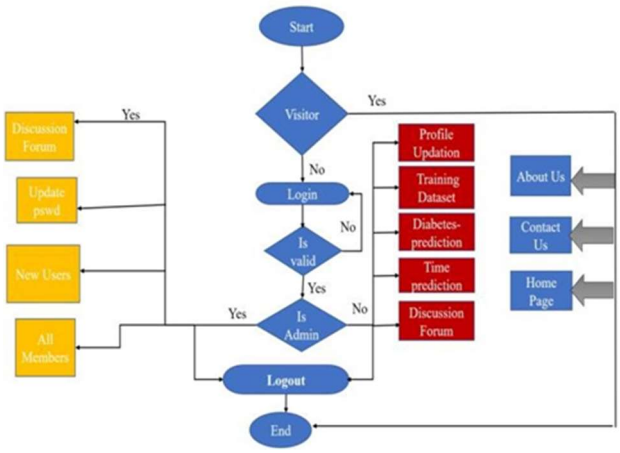

Figure 1: Flow Chart for the Generalized Methodology

In this approach, the Pima Indian data set is used to train a model using a collection of worldwide datasets. Data set contain 21 specifications and approximately 1000 data sets.The dataset's contents and boundaries are:

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

These data are used to train the diabetes prediction framework.

Train Dataset and Test Dataset

The absolute previously set of information used to comprehend the framework is the preparation information. This is the case where we must train the model before setting the feature because the system already has the necessary data. This data is utilized to prepare the machine to function different assignments. It is the knowledge gained by the model as a result of using an algorithm to prepare it and perform errands naturally.

Testing information is used as input by software. It demonstrates how the data changes when the specified module is executed, and is primarily used for testing.

Data Preprocessing

Data preprocessing is a cycle which is utilized to change the crude information with a perfect informational collection. It's the most common way of changing or on the other hand encoding data into a design that a machine can do without much of a stretch parse. The essential capability of information preprocessing in the growing experience is the evacuation of unimportant information and the completion of missing values. in order for machines to be easily trained.

Figure 2: Data Pre-processing.

Extraction of Features

Feature extraction is a method used to alter the important data for outcome features. Quality square is used to decide the attributes of given plans that aid differs among the class of central example subtleties. This technique involves reducing the amount of resources expected to depict a large amount of information. A process of attribute reduction is feature extraction. This is used to expedite and improve the efficiency of regulated learning.

KNN, an ML algorithm

The k-nearest neighbour algorithm, a non-parametric approach developed by Thomas Cover, is used in machine learning for regression and classification. This approach is largely utilised in the sector to classify challenges. The KNN algorithm is a great representation of an occurrence based learning strategy. This method significantly improves accuracy by normalizing training data and using distance to classify objects. The collection of objects for whose classes or object property values are known is used to determine the neighbours. Although there are no explicit training steps needed, it can be considered a training set for the algorithm.

Creating Systems

System design is the technique used to determine the connection point, modules, and information for a framework to fulfill a characterized need. System design is the practice of applying a systems approach. Creating the framework engineering by providing the information and data required for framework execution is the main objective of system design. This project uses a three- tier architecture.

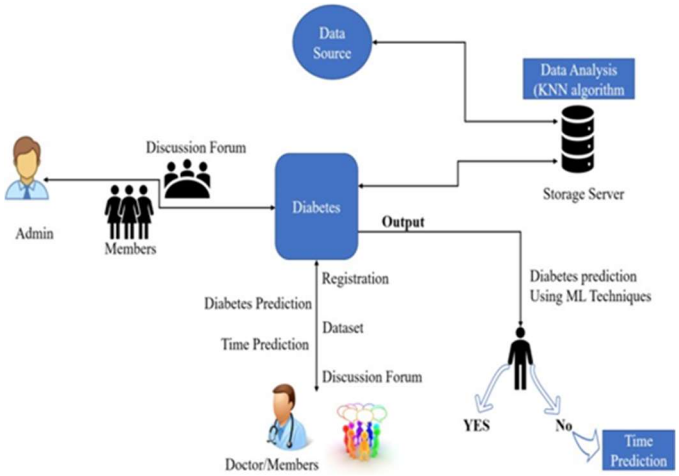

Figure 3 : Architecture Design.

Additionally, one of the classifiers used to categorise the skill has a remarkable exactness of 72.14%. The SMO classifier was used to do that. The Support Vector Machines are constructed using Sequential Minimal Optimization (SMO) (SVM). This is because of the way that an answer for the quadratic programming (QP) improvement issue is expected to make the SVM. It was accomplished via the SMO, which divides the QP into smaller, more manageable issues that may be resolved quickly. Additionally, the amount of memory needed to solve minor issues is kept to a minimum. This enables SMO to manage extraordinarily large datasets. It also features pre-handling capabilities that allow it to substitute for absent qualities and convert seeming characteristics into double ones. Additionally, it automatically normalises the data. This contributes to improving forecast accuracy.

The Bayes Nets produced outstanding outcomes for the Bayes category in a variety of domains, including public safety, scientific research, and aircraft systems. Despite the fact that it works effectively, it is not advised to use it in forecasting problems like the one in this review. This is because the algorithm considers how the factors may affect the outcome. In terms of determining the effects of the components, it outperforms relapse capacities.

The KStar classifier produced results with the lowest accuracy. It is able to handle noisy data and requires less money to prepare the data. Although, with large datasets, its presentation ends up being more effective. Additionally, the parameter k's value must be determined in order to employ this strategy. The cost of calculation is relatively significant because deciding the distance is essential between each occurrence in the training sample.

In conclusion, decision tree algorithms showed the greatest accuracy and are advised for usage in classification and prediction issues.

Various calculations also have a high degree of precision. Therefore, in order to benefit from their assets, we advise including these estimates in the grouping and expectation studies. These computations can also be used to improve the

precision of other Profound or AI processes as well as human- made brainpower methods.

IV. RESULTS & ANALYSIS

Researchers are motivated to test different classifier types and develop new models to expand the exactness of diabetes expectation. This technique uses similar vision to get highly accurate forecasts. Regarding their accuracy and recurrence of purpose, all Machine Learning (ML) and Deep Learning (DL) classifiers developed during the last six years were examined. On the PID dataset, ML classifiers with one or zero recurrence have been carried out to generate recommendations for their application. These machine learning techniques produced accuracy between 68% and 74%. The researchers' greatest accuracy for the DL algorithms was 95%. In order to further increase the precision of predicting the Diabetes condition, the remaining classifiers can be included in a combined model in the future.

Realizing an application, or putting plans, ideas, models, design, and system development into practice, as well as defining the model, standard, and system- or authority-specific algorithms, can

all be referred to as implementation. An implement is defined in computer science as the technological embodiment of an algorithm as a programmed, programming part, or some other kind of PC framework, through PC programming and arrangement. For a specific specification or standard, there may have been many implementations.

V. CONCLUSION & FUTURE WORK

The ability to predict diabetes is crucial in the modern world because of the serious complications it can cause. Because diabetes is the leading cause of death globally. The System model focuses primarily on identifying diabetes using a few parameters. Doctors can predict the onset of diabetes with the help of a framework. so that sufferers can get conventional remedies and therapies. For its prediction, the system used methods like machine learning (ML) in order to generate more precise findings. The diabetic imprint has been the subject of numerous investigations.. For medical clinics and specialists, developing a framework for diabetes illness expectations is beneficial. To help doctors treat patients more effectively, a system forecasts disease in its early stages. The suggested framework is a real-time application developed for several hospitals that can more quickly anticipate sickness. As we employ machine learning algorithms and computations for sickness expectation, we will deliver more precise and advantageous findings..

VI. 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. Deepak Singh Rajput, Saurabh Sharma, Shiv Kumar Tiwari, AK Upadhyay, Ashish Mishra "Medical data security using blockchain and machine learning in cloud computing", Dec. 2020.

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