Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

Improving Healthcare Management: Diabetes Prediction via Extreme Learning Machine

Kunduru Ashwini[1], Golla Chakrapani[2], Gandhavalla Sambasiva Rao[3]

[1,2,3]Assistant Professor, Department of CSE, Malla Reddy Engineering College and Management Sciences, Hyderabad, Telangana.

Abstract

Recently, the health sector is widely adopting artificial intelligence models such as machine learning (ML), deep learning for data analysis, disease prediction, and disease classification. However, the conventional models failed to analyze the data. Therefore, this work is focused on analysis of diabetes prediction using extreme learning machine (DP-ELM) model. Initially, Pima Indian diabetes is considered, which is pre-processed for missing data symbols identification. Then, the statistical features from pre-processed dataset are extracted using principal component analysis (PCA). Then, ELM model is trained with the PCA features and forms the trained feature dataset. Then, a random test combination is applied for ELM testing, which classifies the positive and negative status of diabetes. The simulations proved that; the proposed DP-ELM outperformed in terms of accuracy as compared to existing methods.

Keywords—*Health informatics, Indian Diabetes dataset, Diabetes prediction, Extreme learning machine*

I. INTRODUCTION

In order to function, the human body requires energy. The breakdown of the carbs results in glucose, which is the primary source of fuel for the cells that make up the human body. Insulin is required in order for glucose to be transported into the cells of the body. The pancreas secretes the hormones insulin and glucagon, which are necessary for maintaining normal levels of glucose in the blood. When there is a rise in the glucose level in the blood, the beta cells are activated, and insulin is released into the bloodstream. Insulin is a hormone that allows glucose, which is found in the blood, to enter cells where it may be converted into energy. As a result, there is a tight control over the glucose levels in the blood. Diabetes is a chronic condition that has the potential to produce a crisis in the medical treatment provided all over the globe. The International Diabetes Federation estimates that 382 million individuals throughout the globe are now coping with the effects of diabetes. This number will more than quadruple to 592 million by the year 2035 [1]. However, owing to the complicated interaction of a number of variables, early prediction of diabetes is a job that is particularly difficult for medical professionals to do. Diabetes may impact a variety of human organs, including the kidney, eye, heart, nerves, and feet, among others. Data mining is a method that involves extracting relevant information from massive databases. As there are extremely vast and enormous amounts of data accessible in hospitals and medical-related sectors, data mining is becoming more important. It is a multidisciplinary topic of computer science that incorporates computational process, ML, statistical approaches, classification, clustering, and detecting patterns. Specifically, it is known as pattern discovery. In recent years, approaches from the field of data mining have seen an increased amount of usage in applications such as time-series prediction [2, 3]. There have been a number of different algorithms for data mining that have been suggested for early illness prediction with a greater level of accuracy in order to save human lives and cut down on the expense of treatment [4]. Consequently, making use of these algorithms to make diabetes forecasts have to be done. In the course of our research, we used a total of five distinct supervised learning strategies to carry out our experiment.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

Consequently, the emphasis of this paper is the investigation of the DP-ELM model. Initial consideration is given to pima indian diabetes, which is pre-processed for detection of missing data symbols. Then, PCA is used to extract the statistical characteristics from the pre-processed dataset. The learned feature dataset is then formed by training an ELM model using PCA features. Then, a random test combination is used for ELM testing, which identifies diabetes as positive or negative.

Rest of the article is organized as follows: section 2 describes about the conventional related work with problem statement. Section 3 describes about the process of DP-ELM with mathematical analysis. Section 4 describes about the results and analysis. Section 5 describes the conclusion.

II. LITERATURE SURVEY

The discipline known as health informatics works to create connections between seemingly unrelated concepts on a broad scale. Reading data from dataset linking in software engineering has usually been unsuccessful due to the fact that a healthcare dataset is often found to be both incomplete and noisy. Because it is possible to store data on a massive scale, the field of computer science known as ML is developing at a fast pace. Although there are many ML tools that can be used to analyze data and provide information that may enhance the quality of work for employees as well as physicians, there is presently no approach that can be utilized by developers. The industry offers a large array of data analysis tools, which, when used, may help decision makers uncover fascinating patterns and previously unknown linkages [5]. When attempting to estimate the multivariable exposure-response function, BKMR relied on the R program as a statistical method for gathering information on health impacts [6]. For the purpose of augmentation, Augmentor incorporated the Python image library [7], but for the display of medical treatment plans and patient data, CareVis [8] was used since it was developed specifically for this endeavor. Other applications need the use of COQUITO's graphical user interface [9]. The well-known 3P tools [10] were used for the data analytics of the healthcare industry. Numerous simple applications, such as WEKA, which provided a graphical user interface for a variety of ML algorithms [11], and Apache Spark, which was used for the cluster computing framework [12], are powerful systems that can be used in a variety of applications for the purpose of solving problems by utilizing big data and ML [13]. Software engineering for ML applications (SEMLA) is a conference that examines the difficulties, fresh insights, and practical approaches relating the engineering of ML and artificial engineering [14]. NSGA-II presented methods for use in applications in the real world that make use of more than one objective function in order to improve performance in terms of variety as well as convergence [15]. In clinical genomics, ML algorithms may often be broken down into one of three primary categories: supervised, unsupervised, or semi-supervised [16]. ISRA, which stands for interflow system requirement analysis, was used in order to ascertain the system needs.

In [17] authors proposed random forest algorithm for diabetic’s prediction. However, this method has higher computation complexity. The effectiveness and precision of the various algorithms that were used are analysed and contrasted. The many approaches to ML that were investigated for this research led to the identification of the algorithm that performed the most accurately when predicting diabetes. In [18] authors suggested decision tree algorithm for diabetic’s prediction, which is an extension to random forest. In [19] authors suggested artificial neural network (ANN) based basic networking for diabetic’s prediction, which gains higher classification than traditional ML models. Researchers are getting more interested in diabetes prediction as a field of study in order to train a computer to determine if a patient has diabetes or not by using the appropriate classifier on the dataset being studied. It has been determined, on the basis of work done in the past for research purposes, that the categorization procedure has not been significantly improved. Since of this, a system is necessary because diabetes prediction is a key topic in artificial intelligence models, and this is so that it can

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

manage the challenges that have been found based on past research. In [20] authors utilized support vector machine (SVM) for diabetes prediction. The findings contrasted with other studies that have been done before that employed the same dataset from the body of academic research. It has been shown that the suggested strategy may result in a greater level of accuracy when used to diabetes onset prediction.

III. PROPOSED METHODOLOGY

There are various ML models were developed for prediction of diabetic prediction using artificial intelligence properties. There are variety of ML techniques including Linear SVC, Multinomial Naive Bayes, Random Forest, Logistic Regression, and KNN. They are failed to provide the prediction performance due to low-level feature analysis. Figure 1 shows the proposed DP-ELM model for diabetes prediction. Usually, Indian-pima dataset contain null data, missing data, and non-numeric data, all of which have the potential to reduce the accuracy of ML prediction. In order to get around this issue, this work applies pre-processing operation. Further, the pre-processing operation also converts the dataset into numeric data using sk-learn libraties. Then, the PCA takes a dataset and cleans it up by removing any features that aren't essential before keeping just the most significant ones. This allows for more accurate prediction. This method will first train itself using ELM model through PCA features, after which it will build a train model. This train model will then be used to additional test data in order to do prediction. We are able to teach machines to learn and make predictions without the assistance of humans. Once we have completed the preceding models, we can next apply fresh test data to this model to determine whether or not patient lab results are positive. This work also considered the unit testing module to verify the dependability of the modules that came before it by doing software quality checks, and software verification.

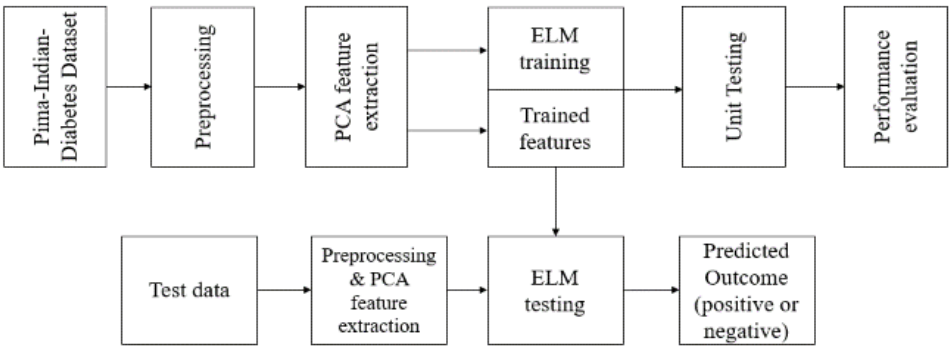

Figure 1. Proposed framework of diabetes prediction.

A. Pre-processing

The raw Indian-pima diabetes dataset contains noises, missing values, which caused to complicated training of ELM model. Further, it will reduce the classification, prediction performance. So, the data preprocessing operation is performed to overcome these problems. The preprocessing operation will replace unknown symbols, missing vales with the known nearest values.

Splitting the Dataset: Our dataset is divided into two distinct categories—the training set and the test set—during the preprocessing phase of ML. These categories are named respectively the training set and the test set. This is one of the most significant activities since enhancing the overall performance of our ML model is one of the keys aims of the processes for data preparation, and this is one of the chores that has to be completed. Take into consideration the following possibility: After training our ML model on a certain dataset, we tested it on a completely other dataset to see how well it performed.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

If this occurs, our model will have a more difficult time understanding the links that exist among the various models. If we train our model incredibly well, and if its training accuracy is also fairly excellent, but then we feed it with a new dataset, then the performance will fall; however, this only happens if we train it exceptionally well. As a result of this, our objective while constructing a model for ML is to make certain that it performs well not just with the training set but also with the test dataset.

B. PCA feature reduction

PCA is a well-known unsupervised learning technique that may decrease the dimensionality of data in many ways. It increases the material's interpretability and decreases the quantity of information that is lost. It makes the data easy to plot in both two and three dimensions and aids in identifying the most significant parts of a dataset. PCA is beneficial for discovering a sequence of linear combinations of the researched variables.

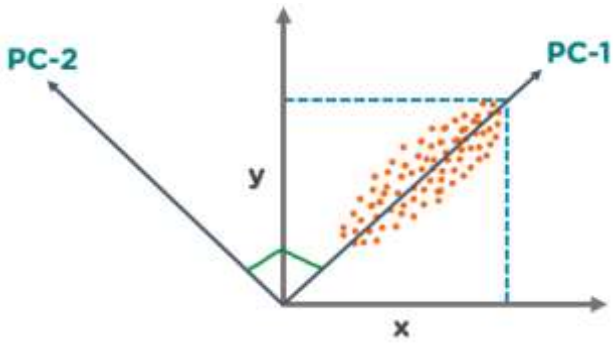

Figure 2. PCA analysis.

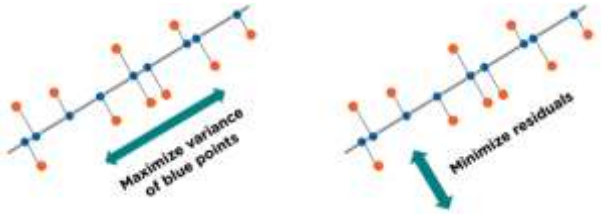

Figure 3. PCA working.

Figure 2 shows the basic process of PCA, which contains the two principal components (PC1) and principal components (PC2). A number of points have been plotted on a two-dimensional. There are two primary elements at play here. The PC1 is the major principal component that is responsible for explaining the greatest amount of variation in the data. Further, figure 3 shows the feature extraction using PCA.

- Step 1: Normalize the data: Before completing the principal component analysis, normalize the data. This will guarantee that each characteristic has a mean value of zero and a variance value of one.

- Step 2: Constructing a square matrix to describe the connection between two or more characteristics in a multidimensional dataset is what is meant by "building the covariance matrix."

- Step 3: Find the Eigenvectors and Eigenvalues: Perform the calculations necessary to determine the eigenvalues and eigenvectors/unit vectors. Scalars known as eigenvalues are used to multiply the covariance matrix's eigenvector in order to get the variance matrix.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

- Step 4: Determine the number of primary components after sorting the eigenvectors in descending order from highest to lowest.

C. ELM Prediction

The ELM algorithm uses the distance between two data points to determine whether or not they are members of a group. K-Nearest Neighbor is the greatest example of this sort of algorithm. The training dataset's real examples are employed as exemplars in the exemplar-based learning algorithm that gave rise to this instance-based learning method. A fresh input vector's proximity to each instance stored is estimated using the distance function by the ELM during generalization, which aids in output class prediction.

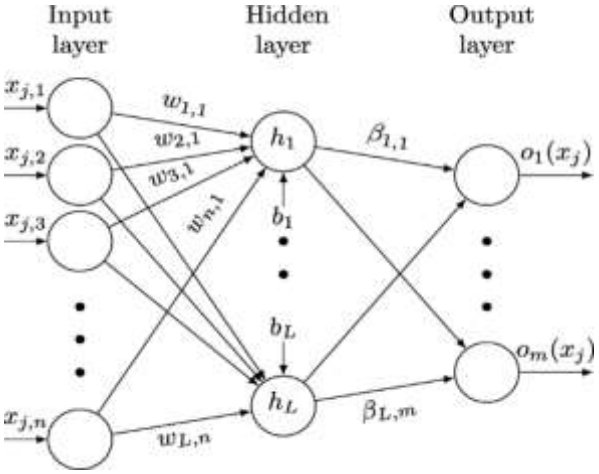

Figure 4. ELM process.

Figure 4 shows the working process of ELM. In order to make judgments based on a set of criteria, ELM uses the data that has been categorised. In this widely used approach, a decision tree is shown with many branches and leaves that represent many elements that might lead to a certain scenario. An ELM is a graph that resembles a tree and specifies the choices and potential consequences.

Although the ELM is a fascinating behavioral machine-learning model and is equivalent to supervised learning, training does not need sample data. Learning is done instead via trial and error. Based on a number of successful results, the optimal suggestion strategy is created for a particular issue. This model can learn how to correspond input and output. As it learns, the algorithm explores random states of action to build a state table, then uses what it has learned to choose the optimum course of action for each state that will allow it to achieve its objective.

D. Unit testing

This work applying classification, clustering, or regression problems, that has been proposed by making use of different sized datasets, and the dataset from Palestine Hospital is being used to implement this concept because that dataset is not available on the internet and also has not been published on the internet; therefore, we are using the dataset from Indian diabetes. We are going to train the ML algorithms listed above using this dataset, and then we are going to do unit testing to ensure that all ML algorithms are providing correct accuracy numbers.

- The subscript for the permeability of vacuum □$_0$, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

IV. RESULTS AND DISCUSSION

This section gives the detailed results analysis of proposed method, which are implemented using DP-ELM model.

A. Dataset

This data collection includes the medical histories of 416 patients with liver disease and 167 patients with other types of liver disease who were gathered from the North East of Andhra Pradesh in India. The "Dataset" column is a class label that is used to classify the groups of people into those who have and do not have liver disease (no disease). This data collection comprises 441 male patient records and 142 female medical records. Figure 5 depicts the dataset screen; all values are taken from the lab report, and the "Class" value is either 0 or 1. The ML algorithm will be trained using the variables from the lab report and the Class Value before generating a model. On the test data below, we will use the generated train model to predict the class label. Since there is no Class label column in the test dataset below, machine learning will predict Class label only based on lab results).

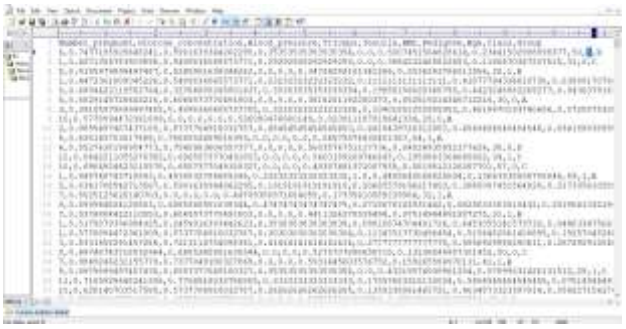

Figure 5. Sample Dataset.

Figure 6 shows the titles of the columns. The numbers in the boxes with minus signs are not significant, only the positive column values are, and the ML algorithm will only train with positive values.

Figure 6. Skewness Matrix.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

In Figure 7, green colour dots are the records which contains no disease and red colour dots are the records which contains disease and this graph generated for all 154 test records. In Figure 8, for each test lab record ML predict whether disease is positive or negative. Figure 9 represents ML algorithm names and y-axis represents accuracy of all those algorithms and from above graph we can conclude that proposed EML is giving better accuracy.

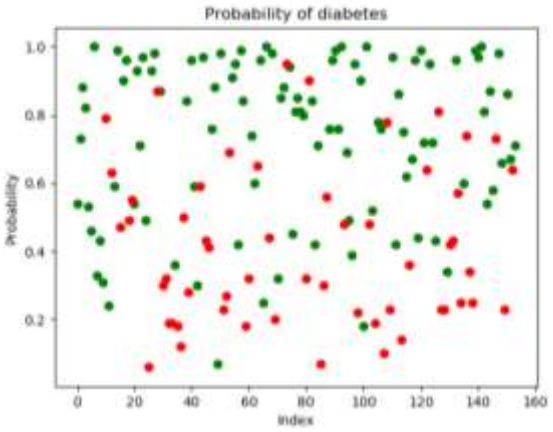

Figure 7. Probability of diabetes.

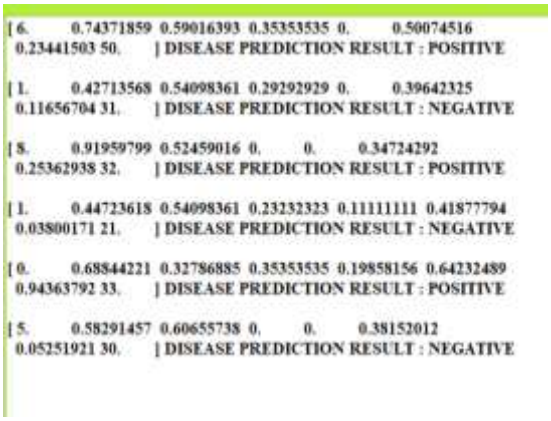

Figure 8. Prediction from test data.

In Table 1, we can see prediction accuracy of each algorithm and from all algorithm's proposed DP-ELM is giving good prediction accuracy and now all ML algorithms. The proposed DP-ELM model resulted in maximum accuracy than conventional classifiers like KNN, Naïve Bayes, Random Forest. Logistic Regression, and Linear SVC.

TABLE I. PERFORMANCE COMPARATION

Model	Accuracy (%)
KNN	63.6363
Naïve Bayes	70.1298
Random	74.6753

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

Forest	
Logistic Regression	75.3246
Linear SVC	59.0909
Proposed DP-ELM	92.8571

V. CONCLUSION

In recent years, the field of medicine has been more interested in using artificial intelligence models like machine learning and deep learning for the purposes of data analysis, illness prediction, and disease categorization. On the other hand, the traditional models were not able to successfully analyze the data. As a result, the examination of the DP-ELM model is the primary emphasis of this paper. In the first, pima indian diabetes is taken into consideration, and then it is pre-processed for the detection of missing data symbols. The principal component analysis (PCA) is then used to extract the statistical characteristics from the pre-processed dataset. The next step is to train the ELM model using the PCA features, which results in the formation of the trained feature dataset. After that, an arbitrary test combination is used for ELM testing, which determines whether or not a person has diabetes based on their positive or negative status. The simulations demonstrated that the suggested DP-ELM performed better in terms of accuracy when compared to the approaches that are already in use. This work is extended with advanced artificial intelligence models for better performance.

REFERENCES

[1] Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[2] S. V. N. Sreenivasu, S. Gomathi, M. Jogendra Kumar, Lavanya Prathap, Abhishek Madduri, Khalid M. A. Almutairi, Wadi B. Alonazi, D. Kali, S. Arockia Jayadhas, "Dense Convolutional Neural Network for Detection of Cancer from CT Images", BioMed Research International, vol. 2022, Article ID 1293548, 8 pages, 2022. https://doi.org/10.1155/2022/1293548

[3] Kotturi, S.H.K., Sreenivasu, S.V.N. "Detection of Pneumonia Using Convolution Neural Networks," In: Shakya, S., Du, KL., Haoxiang, W. (eds) Proceedings of Second International Conference on Sustainable Expert Systems . Lecture Notes in Networks and Systems, vol 351. Springer, Singapore, 2022. https://doi.org/10.1007/978-981-16-7657-4_19

[4] Emoto, Takuo, et al. "Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease." Heart and vessels 32.1 (2017): 39-46.

[5] Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform." Knowledge-Based Systems 37 (2013): 274-282.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

[6] Fatima, Meherwar, and Maruf Pasha. "Survey of ML Algorithms for Disease Diagnostic." Journal of Intelligent Learning Systems and Applications 9.01 (2017): 1.

[7] Singh, Harimohan, and Amit Kumar Gupta. "Implementation and Application of Machine Learning in Health Care: A Review." Proceedings of the Second International Conference on Information Management and Machine Intelligence. Springer, Singapore, 2021.

[8] Johri, Prashant, Vivek sen Saxena, and Avneesh Kumar. "Rummage of machine learning algorithms in cancer diagnosis." International Journal of E-Health and Medical Communications (IJEHMC) 12.1 (2021): 1-15.

[9] Tiwari, Mukesh, Jan Adamowski, and Kazimierz Adamowski. "Water demand forecasting using extreme learning machines." Journal of Water and Land Development 28.1 (2016): 37-52.

[10] U-;;ar, AyegUI, Yakup Demir, and CUneyt GUzeli. "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering." Neural Computing and Applications 27.1 (2016): 131- 142.

[11] Boyd, C. R.; Tolson, M. A.; Copes, W. S. (1987). "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score". The Journal of trauma. 27 (4): 370 - 378. doi: I 0.1097/00005373-198704000-00005. PMID 3106646.

[12] Kologlu M., Elker D., Altun H., Sayek I. Validation of MPI and OIA II in two different groups of patients with secondary peritonitis II Hepato-Gastroenterology. - 2001. - Vol. 48, N2 37. - pp. 147-151

[13] Kologlu M., Elker D., Altun H., Sayek 1. Validation of MPI and OIA II in two different groups of patients with secondary peritonitis II Hepato-Gastroenterology. - 2001. - Vol. 48, N2 37. - pp. 147-151

[14] Laura Aurialand Rouslan A. Moro2, "Support Vector Machines (SVM) as a Technique for Solvency Analysis " .Symp. Computational Intelligence in Scheduling (SCIS 07), ASME Press, Dec. 2007, pp. 57-64, doi: 1 0.11 09/SCIS.2007.357670.

[15] Zissis, Dimitrios (October 2015). "A cloud based architecture capable of perceiving and predicting multiple vessel behaviour". Applied Soft Computing. 35: 652-661. doi:10.1016/j.asoc.2015.07.002.

[16] Graves, Alex; and Schmidhuber, JUrgen; Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks, in 1010 Bengio, Yoshua; Schuurmans, Dale; Lafferty, John; Williams, Chris K. /.; and Culotta, Aron (eds.), Advances in Neural Information Processing Systems 22 (NlPS'22), December 7th-10th, 2009, Vancouver, BC, Neural Information Processing Systems (NIPS) Foundation, 2009, pp. 545-552.

[17] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.

[18] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of ML Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.

ISSN: 0975-3583, 0976-2833 VOL13, ISSUE 09, 2022

[19] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using ML Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13

[20] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.