Enhancing Health Data Prediction with Software Engineering and Machine Learning: An Application for Health Systems

Kondragunta Rama Krishnaiah, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: kondraguntark@gmail.com

Alahari Hanumant Prasad, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: hanuma.alahari@gmail.com

Abstract

In recent times, machine learning has garnered significant attention as a cutting-edge area of research. As a result, our study focuses on exploring the fascinating intersection between software engineering and machine learning within the realm of health systems. We have introduced an innovative framework dedicated to health informatics. This framework is structured around four crucial modules: software, machine learning, machine learning algorithms, and health informatics data. By utilizing the proposed methodology, we have effectively organized tasks within this framework. The primary goal is to provide researchers and developers with a fresh perspective on health informatics software, incorporating engineering principles. As a result, developers are equipped with a comprehensive roadmap to design health applications, complete with system functions and software implementations. To power the proposed approach, we employ principal component analysis (PCA) for feature extraction and reduction. This technique plays a pivotal role in simplifying complex datasets and facilitating efficient data analysis. Additionally, the proposed model leverages the extreme learning machine (ELM) for prediction problems, contributing to the accurate forecasting of health-related outcomes. In our experimentation, we employed the Indian Diabetes dataset to conduct simulations. Our proposed ELM demonstrated exceptional performance, surpassing the state-of-the-art approaches in terms of predictive accuracy and efficiency.

Keywords: Software engineering, Machine learning, health informatic, Indian Diabetes dataset, extreme learning machine.

1. INTRODUCTION

Human body needs energy for activation. The carbohydrates are broken down to glucose, which is the important energy source for human body cells. Insulin is needed to transport the glucose into body cells. The blood glucose is supplied with insulin and glucagon hormones produced by pancreas. Insulin hormones produced by the beta cells of the islets of Langerhans and glucagon hormones are produced by the alpha cells of the islets of Langerhans in the pancreas. When the blood glucose increases, beta cells are stimulated, and insulin is given to the blood. Insulin enables blood glucose to get into the cells and this glucose is used for energy. So, blood glucose is kept in a narrow range. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million [1]. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors. Diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data mining is a process to extract useful information from large database, as there are very large and enormous data available in hospitals and medical related diabetes. It is a multidisciplinary field of computer science which involves computational process, machine learning, statistical techniques, classification, clustering and discovering patterns. Recently, Data mining techniques have been widely used in predicting the data like time-series [2, 3]. A number of data mining algorithms have been proposed for early prediction of disease with higher accuracy in order to save human life and reduce the treatment cost [4]. Thus, applying these algorithms to predict diabetes should be done. In our work, we used five different supervised learning methods to conduct our experiment.

The field of health informatics (HI) aims to provide a largescale linkage among disparate ideas. Normally, a healthcare dataset is found to be incomplete and noisy; as a result, reading data from dataset linkage traditionally fails within the discipline of software engineering. Machine learning (ML) is a rapidly maturing branch of computer science since it can store data on a large scale. Many ML tools can be used to analyze data and yield knowledge that can improve the quality of work for both staff and doctors; however, for developers, there is currently no methodology that can be used. Regarding software engineering, there has been a lack of approaches to evaluating which software engineering tasks are better performed by automation and which require human involvement or human-in-the-loop approaches [1]. Big data has many challenges regarding analysis challenges for real-world big data [2], including OLAP mass data, mass data protection, mass data survey and mass data dissemination. Recently, a set of frameworks have been used to develop data analysis tools such as Win-CASE [3] and SAM [4]. The market has vast data analysis tools that can discover interesting patterns and hidden relationships to support decision makers [5]. BKMR used the R package as a statistical approach on health effects to estimate the multivariable exposure-response function [6]. Augmentor included the Python image library for augmentation [7], while for the visualization of medical treatment plans and patient data, CareVis was used [8], as it was designed for this task. Other applications require a visual interface using COQUITO [9]. For health-care data analytics, the widely known 3P tools [10] were used. Many simple applications, such as WEKA, which provided a GUI for many machine learning algorithms [11], while Apache Spark was used for the cluster computing framework [12], are powerful systems that can used in various applications for solving problems using big data and machine learning [13]. Software engineering for machine learning applications (SEMLA) discusses the challenges, new insights, and practical ideas regarding the engineering of ML and artificial engineering (AI) [14]. NSGA-II proposed algorithms for real-world applications that include more than one objective function for enhancing performance in terms of both diversity and convergence [15]. ML algorithms in clinical genomics generally come in three main forms: supervised, unsupervised and semi-supervised [16]. Interflow system requirement analysis (ISRA) has been used to determine the system requirements.

2. Literature survey

K.Vijiya Kumar et al. [17] proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in ma- chine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Muhammad Azeem Sarwar et al. [18] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research. Tejas N. Joshi et al. [19] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project pro- poses an effective technique for earlier detection of the diabetes disease.

Nonso Nnamoko et al. [20] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with simi- lar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Deeraj Shetty et al. [21] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patients database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patients database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

3. Proposed Methodology

In proposed system the combining Software Engineering and Machine Learning algorithms to improve disease prediction in health care systems and to minimize time taken to predict disease as we don’t have enough hospitals or bed to accommodate growing number of patients and we can solve this problem of predicting disease with less time by employing software and machine learning algorithms. Proposed method concept is known as SEMLHI Software Engineering with Machine Learning for Health Data).

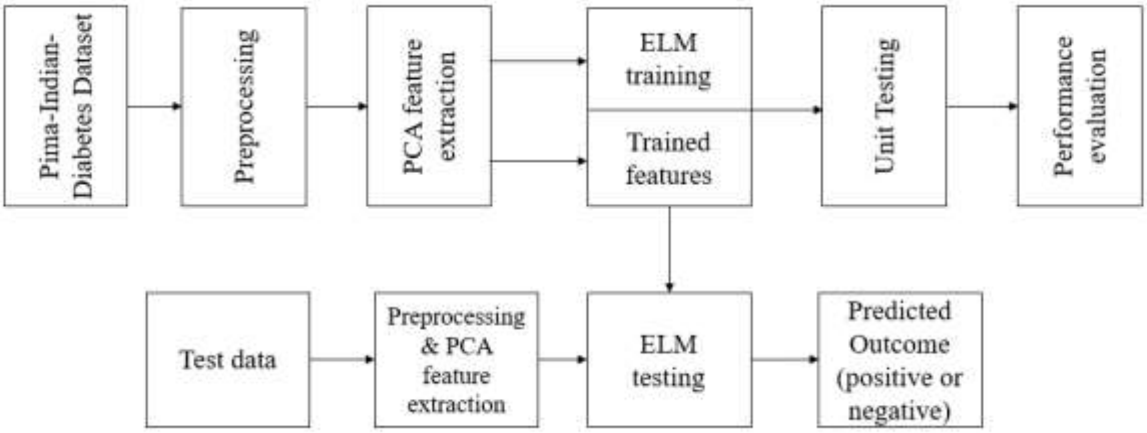


Figure 1. Proposed framework.

Figure 1 shows the proposed framework. Propose SEMLHI consists of 4 components. In proposed work by using various size of dataset we are applying classification, clustering ore regression and to implement this concept is using Palestine Hospital dataset and this dataset not available on internet and also not publish this dataset on internet so using INDIAN DIABETES dataset. we will use this dataset to train above ML algorithms and then perform UNITTESTING to check all ML algorithms are giving accurate accuracy values.

3.1 Preprocessing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data preprocessing task.

Need of Data Preprocessing: A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

3.2 Splitting the Dataset

In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset.

3.3 ELM Prediction

ELM is a kind of advanced neural network, consists of three layers such as input layer, hidden layer (number of neurons) and an output layer. The input layer captures the input variable, hidden layers make a linear relationship among the variables and the output layer presents the predicted value. The following principle that differentiates ELM from other traditional NN is based on the parameters of the feed-forward network, inputs weights and biases provided to the hidden layer. In ELM, the bias of the hidden layer and input weight are randomly generated, and the output is calculated by the Moore–Penrose generalized inverse of the hidden layer output matrix. The randomly chosen input weight and hidden layer biases learn the training samples with minimum error. After randomly choosing the input weights and the hidden layer biases, SLFNs can be simply considered as a linear system. The main advantage of ELM, its structure does not depend on network parameters which produce stability. Hence it is useful for classification, regression, and clustering. Therefore, we adopted ELM as a classification model in predicting the software quality. Figure 2 shows the architecture of ELM with four input layers, ten hidden layers, and three output layers. The process of training and testing the ELM contains a network with two vectors of input vector and target output vector.

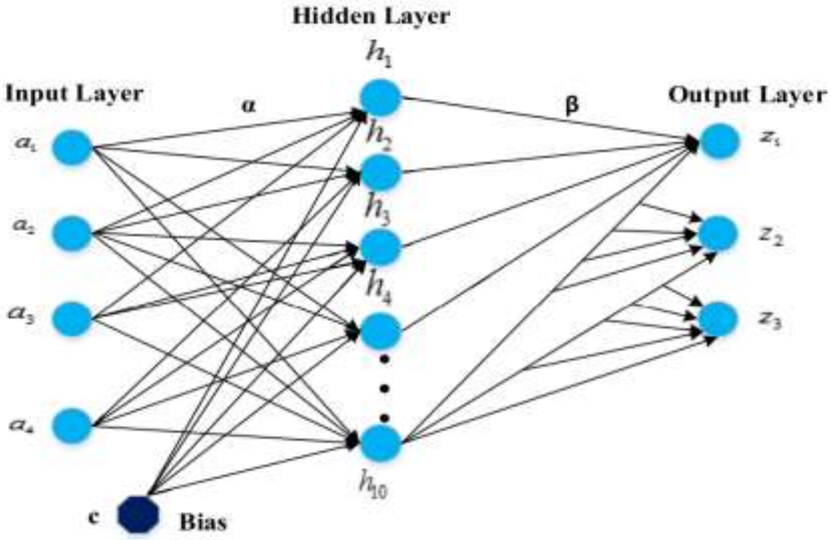


Figure 2. The architecture of the extreme learning machine

4. Results

In dataset, all values are the lab report values and 'Class' value contains 0 or 1 and ML algorithm will train with lab report values and Class Value and then generate a model. Generated train model we will apply on below test data to predict class label.

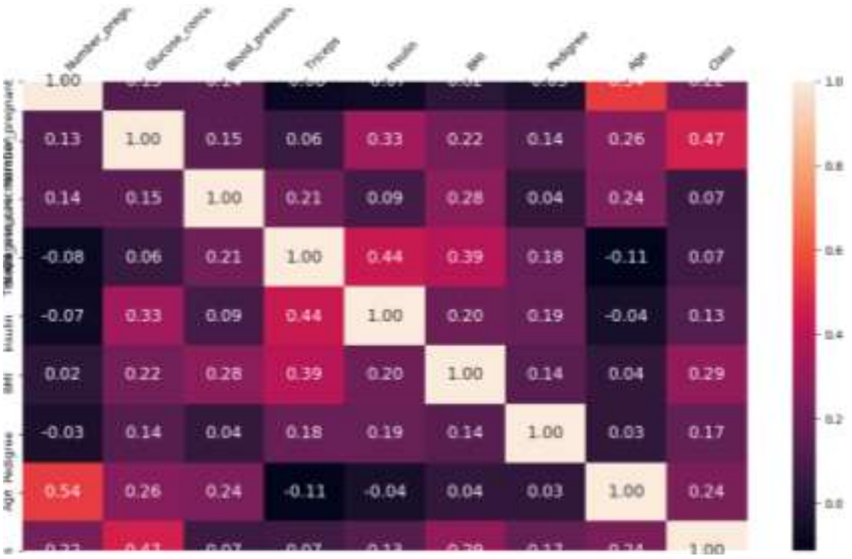


Figure 3. Skewness Matrix

In Figure 3, we can see names of columns and in boxes values with minus symbols are not important and only positive column values are important and ML algorithm will train only with positive values. In Figure 10, 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. Now close above graph to see all ML prediction accuracy

Table 1. Performance comparation

Model	KNN	Naïve	Random	Logistic	Linear SVC	Proposed

		Bayes	Forest	Regression		ELM
Accuracy	63.6363	70.1298	74.6753	75.3246	59.0909	92.8571

In Table 1, we can see prediction accuracy of each algorithm and from all algorithm's proposed ELM is giving good prediction accuracy and now all ML algorithms.

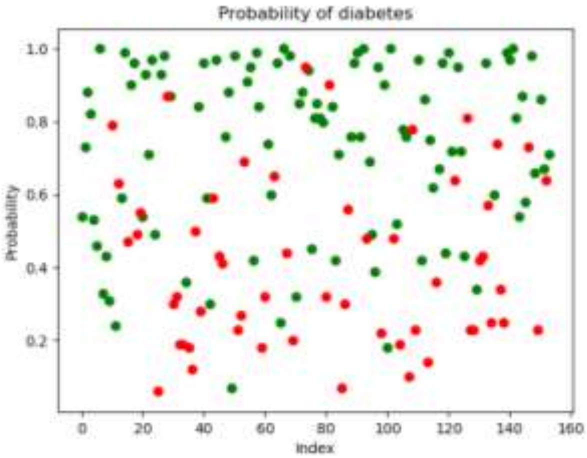
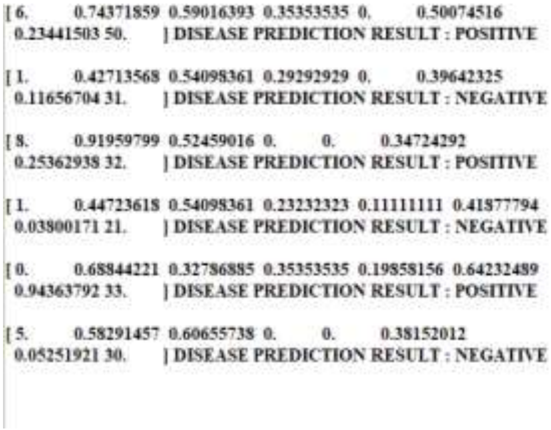


Figure 4. Probability of diabetes



```
[ 6.         0.74371859  0.59016393  0.35353535  0.         0.50074516
 0.23441503 50.        ] DISEASE PREDICTION RESULT : POSITIVE

[ 1.         0.42713568  0.54098361  0.29292929  0.         0.39642325
 0.11656704 31.        ] DISEASE PREDICTION RESULT : NEGATIVE

[ 8.         0.91959799  0.52459016  0.         0.         0.34724292
 0.25362938 32.        ] DISEASE PREDICTION RESULT : POSITIVE

[ 1.         0.44723618  0.54098361  0.23232323  0.11111111  0.41877794
 0.03800171 21.        ] DISEASE PREDICTION RESULT : NEGATIVE

[ 0.         0.68844221  0.32786885  0.35353535  0.19858156  0.64232489
 0.94363792 33.        ] DISEASE PREDICTION RESULT : POSITIVE

[ 5.         0.58291457  0.60655738  0.         0.         0.38152012
 0.05251921 30.        ] DISEASE PREDICTION RESULT : NEGATIVE
```

Figure 5. Prediction from test data

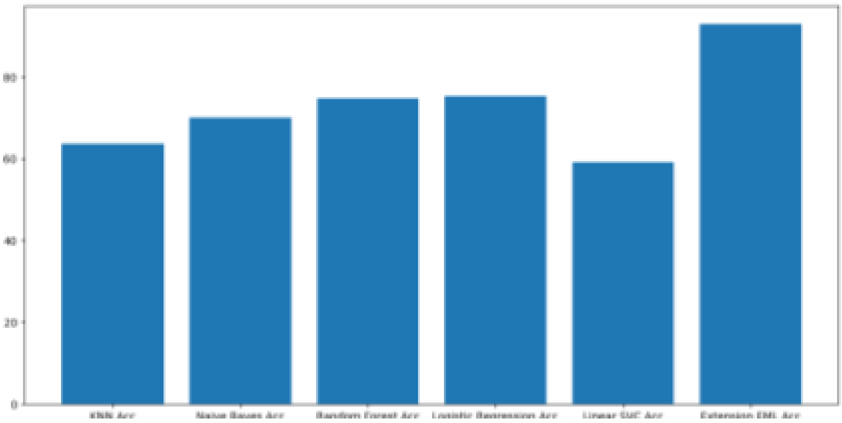


Figure 6. Graphical Representation of comparation

In Figure 5 for each test lab record ML predict whether disease is positive or negative. Figure 6 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.

5. Conclusion

This research introduced a new methodology, that can develop health informatics application using machine learning. Our methodology used the grounded theory methodology to develop SEMLHI framework. Developers use SEMLHI methodology to analyse and developing software for the HI model and create a space in which SE and ML experts could work on the ML model lifecycle. Proposed framework includes a theoretical framework to support research and design activities that incorporates existing knowledge. Our work introduces a new approach form clustering and classification for ML in HI. SEMLHI methodology includes seven-phase, designing (encode data and Define outlier and cleaning up the data), implementing (Verification & Validation), maintaining and defined Workflows, structured Information, security and privacy, testing and performance, and reusing software applications. SEMLHI framework includes four modules that organize the tasks for each module and introduce a SEMLHI Methodological that enable researchers and developer to analyze health informatics software from an engineering perspective. The ultimate goal from a SEMLHI Methodological is to define a standardized methodology for software development in the Health area and include all stages from defining the problem until developing the application and get the result with the test stage.

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] Berry, Michael 1., and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997.

[3] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[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.

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

[7] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[8] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[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 Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.

[19] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) January 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.

[21] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.