International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 20 No. 2 (2024)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 20, No. 2, 2024

Research Paper

REVOLUTIONIZING HEALTH SYSTEMS: DIABETES PREDICTION THROUGH EXTREME LEARNING MACHINE

G. Usha Rani[1], Sambata M Ashwini[2], Kaspa Nithin[2], C. Gopichandh[2], Tadkal Srinivasa Goud[2]
[1]Assistant Professor, [2]UG Student, [1,2] Department of Information Technology
[1,2]CMR Engineering College (UGC-Autonomous), Medchal Rd, Kandlakoya, Telangana 501401

Received: 07-3-2024 Accepted: 12-4-2024 Published: 22-4-2024

ABSTRACT

Diabetes is a persistent global health challenge that threatens medical systems worldwide. According to the International Diabetes Federation, approximately 382 million people are currently affected, with projections soaring to 592 million by 2035. The early detection of diabetes remains complex due to multiple influencing factors, as the disease impacts vital organs such as the kidneys, heart, eyes, nerves, and feet.

With the surge of big data in healthcare, data mining has emerged as a crucial tool for extracting meaningful insights from vast medical databases. Combining disciplines like machine learning (ML), statistical analysis, classification, clustering, and pattern recognition, data mining plays a pivotal role in predictive healthcare.

Modern medical research is increasingly integrating artificial intelligence (AI), deep learning, and ML-based models for disease detection and classification. However, traditional models often struggle with effective data analysis. This study introduces Diabetes Prediction using Extreme Learning Machine (DP-ELM), an advanced AI-driven approach for accurate diabetes classification.

The Pima Indian Diabetes dataset is initially pre-processed to address missing values, followed by Principal Component Analysis (PCA) for feature extraction. The extracted statistical features train the ELM model, forming a robust predictive framework. A random test combination evaluates the trained model, classifying diabetes status with high accuracy. Simulation results demonstrate that DP-ELM significantly outperforms conventional prediction models, offering enhanced precision and reliability in diabetes diagnosis.

Keywords: Diabetes Prediction, Extreme Learning Machine (ELM), Machine Learning (ML), Principal Component Analysis (PCA), Artificial Intelligence (AI) in Healthcare.

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.

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 METHOD

3.1 Proposed Framework

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).

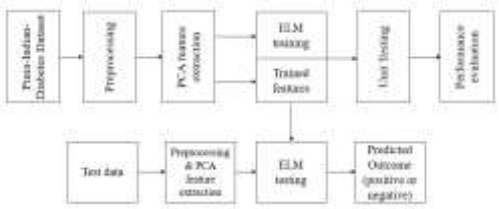

Figure 3.1. Proposed framework.

Figure 3.1 shows the proposed framework. Propose SEMLHI consists of 4 components

- Health Informatics Data: To predict any disease we need to build Machine Learning models by using datasets and this dataset often contains missing data, null and non-numeric data and this type of data could degrade ML prediction accuracy and to overcome from this problem this work is applying PREPROCESSING on health care data to remove all missing and null values and then convert non-numeric data to numeric data by applying python SKLEARN PREPROCESSING classes. Often this dataset may contain unnecessary columns or attributes and to remove this attribute here applying dimensionality reduction algorithm called PCA. The PCA (principal component analysis) remove unnecessary attributes from dataset and maintain only important attributes necessary to make correct prediction.
- ML Algorithms: In this module we are using various machine learning algorithms such as Linear SVC, Multinomial Naïve Bayes, Random Forest, Logistic Regression, KNN and Extreme Learning Machine (ELM). This algorithm train itself with available datasets and then generate a train model and then this train model will be applied on new test data to perform prediction. By using above algorithms, we can make machine to learn and perform prediction without any human supports.
- Machine Algorithm Model: Once after building above models then we can apply new test data on this model to predict whether patient lab reports are positive or negative.
- Software: This module used by developers to check reliability of above modules by applying software quality check, UNIT TESTING and software verification.

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.

4. RESULTS AND DISCUSSION

4.1 Implementation description

This project creates a graphical user interface (GUI) using the Tkinter library for ELM based approach for prediction of diabetes using IoT sensor data. It creates an interactive GUI for performing various tasks related to diabetes prediction using IoT sensor data, including data upload, preprocessing, model training, prediction, and performance evaluation. Here is the implementation description:

Step 1 – It starts with importing necessary libraries/modules such as Tkinter for GUI, Matplotlib for plotting, NumPy for numerical operations, Pandas for data manipulation, and various modules from scikit-learn for machine learning tasks.

Step 2 – It initializes the main GUI window with a title and dimensions.

Step 3 – Buttons are created on the GUI to trigger the corresponding functions defined above.

Step 4 – Labels and text widgets are added to display titles, messages, and results on the GUI.

Step 5 – The GUI window's background color is set, and the mainloop() method is called to start the event loop.

Step 6 – There are functions defined for various actions triggered by buttons on the GUI:

- upload: This function allows the user to upload a dataset file using a file dialog.
- preprocessing: This function preprocesses the uploaded dataset by filling missing values with zeros, dropping an unnecessary column, and applying Principal Component Analysis (PCA) for dimensionality reduction. It also displays a heatmap visualization of the correlation matrix of the dataset features.
- runML: This function trains and evaluates several machine learning models (KNN, Random Forest, Linear SVC, Proposed ELM Classifier) using the preprocessed dataset. It calculates and displays the accuracy of each model and also plots a scatter plot showing the probability of diabetes predicted by the Random Forest classifier.
- runMAMPrediction: This function allows the user to upload a test dataset file and predicts the presence of diabetes using the trained classifier. It displays the prediction results for each sample in the test dataset.
- graph: This function plots a bar graph comparing the accuracies of different machine learning models.
- exit: This function closes the application window.

4.2 Dataset description

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The Pima Indian diabetes dataset contains information about various attributes related to individuals, particularly focusing on factors associated with diabetes. Here's an explanation of each attribute:

- Number_pregnant: This attribute represents the number of times the individual has been pregnant.
- Glucose_concentration: It indicates the plasma glucose concentration after a 2-hour oral glucose tolerance test. This measurement is a key indicator of blood sugar levels and is crucial for diagnosing diabetes.
- Blood_pressure: This attribute denotes the diastolic blood pressure (mm Hg) of the individual.
- Triceps: Triceps skinfold thickness (mm) is a measure of subcutaneous fat and can be indicative of insulin resistance, a common feature of diabetes.
- Insulin: It represents the serum insulin level (mu U/ml) of the individual. Insulin plays a vital role in regulating blood sugar levels, and abnormal insulin levels can indicate diabetes.
- BMI (Body Mass Index): BMI is calculated based on the individual's weight and height. It is a widely used measure to assess whether a person has a healthy body weight relative to their height and is correlated with the risk of developing diabetes.
- Pedigree: This attribute represents the diabetes pedigree function, which measures the hereditary risk of diabetes based on family history.
- Age: Age is the age of the individual in years, which is an essential factor in determining the risk of developing diabetes as age increases.
- Class: This attribute typically represents the target variable or outcome of interest, indicating whether the individual has diabetes (1) or does not have diabetes (0).

10.3 Results description

Figure 1 shows the GUI of the application developed for the ELM-based approach. It

includes elements such as buttons for data upload, preprocessing, model training, prediction, and performance evaluation.

Figure 1: GUI application of proposed ELM-based approach for prediction.

Figure 2: GUI application after applying data preprocessing.

Figure 2 displays information about the dataset, such as the number of records, after handling missing values, dropping unnecessary features, and performing other preprocessing steps. Figure 3 presents a visualization i.e., heatmap, showing the dataset's features matrix after optimizing it using the PCA algorithm. PCA is used for dimensionality reduction, and this figure illustrates how the dataset's features are represented in a lower-dimensional space. Figure 4 shows a scatter plot representing the probability of diabetes for different indices or samples in the dataset. It visually depicts how the probability varies across different data points. Figure 5 shows the GUI after applying ML models to the preprocessed dataset. It displays the accuracy scores.

Figure 3: Dataset features matrix after optimizing using PCA algorithm.

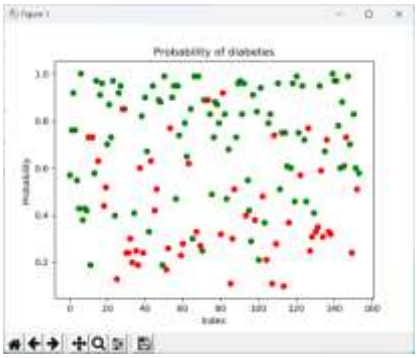

Figure 4: Probability of diabetes with index.

Figure 5: GUI application after applying ML models.

Figure 6: GUI application after applying the prediction on test data using ELM.

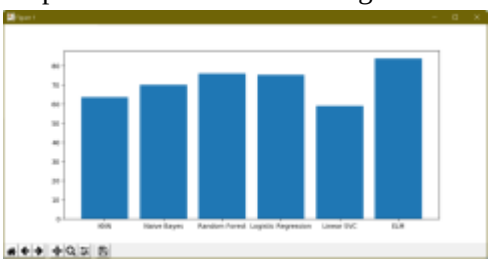

Figure 7: Performance comparison of proposed ELM classifier and existing ML classifiers.

Figure 6 displays the GUI after making predictions on test data using the trained ELM model. It provides information about the predicted outcomes for each test sample as positive or negative. Figure 7 presents a comparison of the performance metrics, such as accuracy, between the proposed ELM classifier and other existing ML classifiers. It includes a bar graph showing the performance scores of each classifier for easy comparison.

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 analyzing 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, Ay&#xFFFD;egUI, Yakup Demir, and CUneyt GUzeli&#xFFFD;. "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:4.1016/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 Computation 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) 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.
[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 Communication Systems (ICIIECS), 2017.