Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, No. 1, July : 2023

HYBRID ENSEMBLE MACHINE LEARNING MODELS FOR ACCURATE DIABETES PREDICTION USING MULTILAYER PERCEPTRON WITH EXTREME MACHINE LEARNING AND SUPPORT VECTOR CLASSIFICATION

Swati, M.tech Scholar, Department of Computer Science and Engineering, IITM Murthal, Haryana
Preeti Nehra, Assistant Professor, Department of Computer Science and Engineering, IITM Murthal, Haryana

Abstract
The prompt diagnosis and effective treatment of diabetes rely heavily on precise prognosis of the disease. In this study, we combine Multilayer Perceptron (MLP), Extreme Machine Learning (EML), and Support Vector Classification (SVC) to provide a novel ensemble machine learning approach for diabetes prediction. The proposed method has the potential to increase prediction accuracy by drawing on the advantages of a number of different techniques. The MLP model is able to faithfully capture the complex non-linear relationships inherent in the data, while the EML model relies on a simplified learning framework to improve its computational efficiency. To top it all off, the SVC model effectively exploits the potential to manage high-dimensional data and nonlinear decision limits. The hybrid ensemble model is trained and tested on a large health database. Important demographic and clinical data may be found in this repository. The experimental findings demonstrate that the proposed strategy delivers superior accuracy in diabetes prediction compared to both standalone models and state-of-the-art machine learning methods. Successfully combining the complementary qualities of MLP, EML, and SVC in a prediction model for diabetes yields a robust and reliable model..
Keywords: MLP, EML, SVM, ML, Diabetes, Prediction
.

I. Introduction

Diabetes is a metabolic illness that is chronic and is defined by high blood sugar levels. Diabetes may either be caused by an inadequate production of insulin (in the case of type 1 diabetes) or an inability of the body to efficiently utilise insulin (in the case of type 2 diabetes) [1]. It is a major public health issue that affects millions of people all over the globe and may result in a variety of consequences, such as cardiovascular illnesses, renal difficulties, and impaired eyesight [2]. Diabetes is a condition that must be well managed in order to reduce the risk of complications and enhance patient outcomes [3]. One of the most important aspects of this care is the early identification and accurate prediction of diabetes. In recent years, machine learning algorithms have developed as useful tools for the prediction and detection of illness. These algorithms make use of the massive volumes of health data that are already accessible. These algorithms are capable of properly analyzing the intricate links and patterns that exist within the data, which provides useful insights for predictive modeling [4]. In particular, hybrid machine learning models, which incorporate numerous methods into a single model, have showed promise in boosting both accuracy and resilience in illness prediction tasks [5].

It is possible to further improve the accuracy and reliability of diabetes prediction by developing a hybrid machine learning model for diabetes prediction utilizing a mix of Multilayer Perceptron (MLP), Extreme Machine Learning (EML), and Support Vector Classification (SVC) [6]. [7] The multilayer perceptron (MLP) is an artificial neural network that is capable of capturing complicated non-linear correlations present in the data. EML provides a more straightforward learning paradigm that results in increased computational efficacy [8]. On the other hand, SVC is a robust algorithm that is able to successfully manage high-dimensional data as well as non-linear decision boundaries [9].

The hybrid model may utilize the benefits of these algorithms by combining their capabilities, which results in a more accurate and robust prediction model [10]. For example, the hybrid model can use the capacity of MLP to capture complicated connections, the computing efficiency of EML, and the handling of non-linear decision boundaries by SVC.
In addition, using Principal Component Analysis (PCA) in the model may be helpful in both lowering the size of the feature space and raising the level of computing efficiency [11]. PCA takes the original characteristics and converts them into a new collection of variables that are called principle components. These variables are not connected with each other, and they capture the greatest amount of variation in the data [12]. This method of reducing the dimensionality of data may assist in the removal of characteristics that are unnecessary or redundant, hence improving the overall performance of the model [13].
The purpose of this investigation is to create a hybrid machine learning model for predicting diabetes by using MLP, EML, and SVC, as well as including PCA into the learning process. Using the strengths of these algorithms and the dimensionality reduction capabilities of PCA [14], the model seeks to attain a high level of accuracy in its predictions about the presence or absence of diabetes. The performance of the model will be analyzed together with its training and evaluation using a large health database, and it will be compared to the performance of other machine learning methodologies [15]. In conclusion, the creation of a hybrid machine learning model for the prediction of diabetes with MLP, EML, and SVC with the addition of PCA has a significant amount of promise for accurate and dependable disease prediction. The model intends to contribute to better patient outcomes, early identification, and effective treatment of diabetes by leveraging the power of these algorithms and adding strategies for dimensionality reduction.

II. Implementation

The following are the processes that make up the approach for constructing the hybrid machine learning model for diabetes prediction using the Pima health dataset and including Principal Component Analysis (PCA). This model will be used to make predictions about diabetes.

The Pima health dataset, which include pertinent clinical and demographic information, is retrieved and checked for any missing values or outliers. This is the first step in the data preprocessing process. Imputation methods such as the mean imputation and regression imputation may be used to fill in the blanks left by missing data. It is possible to deal with outliers by either deleting them or replacing them with values that are more acceptable. In addition, the dataset is separated into the feature matrix (X) and the goal variable (y), where the presence or absence of diabetes is represented by the target variable.

PCA, or principal component analysis, is done in order to decrease the number of dimensions included inside the feature matrix. PCA is a method for reducing the dimensionality of data by transforming the initial characteristics into a new collection of variables called principle components, which are not connected with each other. These components account for the greatest amount of the data's inherent variability. The dimensionality of the feature matrix may be lowered while maintaining the information that is most vital to understanding the problem at hand if only those primary components are kept which can account for a significant amount of the variation.

In order to construct the hybrid machine learning model, the data must first be preprocessed and then subjected to principal component analysis. The MLP, EML, and SVC algorithms have been combined in this model so that it may make the most of each of their respective capabilities. The MLP technique, which is based on neural networks, is able to capture complicated non-linear correlations present within the data. EML, on the other hand, provides a learning framework that is easier to understand and increases the effectiveness of computational operations. A sophisticated

classification method known as SVC is capable of handling high-dimensional data as well as nonlinear decision boundaries.

Training and Assessment: The preprocessed and reduced feature matrix, in addition to the target variable, is partitioned into training and testing sets. After that, the MLP, EML, and SVC training algorithms are applied to the training set in order to train the hybrid model. In order to get the best possible performance, the model parameters are fine-tuned using methods such as grid search and random search. The trained hybrid model is then used to generate predictions on the testing set, after which evaluation metrics such as accuracy, precision, recall, and F1 score are computed. This allows the model to be evaluated. In order to determine whether or not the hybrid model is superior than the individual MLP, EML, and SVC models, in addition to other baseline models, the performance of the hybrid model is evaluated and compared.

Analysis and interpretation of the findings the outcomes of the hybrid model, including accuracy and performance measures, are studied and explained here. It is possible to carry out a feature importance analysis in order to get an understanding of the factors that have the most important effect on the forecasts. This study contributes to the process of explaining how the model arrived at its conclusions and offers new perspectives on the components that play a role in diabetes prediction. It is essential that the model be interpretable for medical professionals so that they can comprehend and have faith in the predictions made by the model when applied to actual-life situations in the medical field.

Overall, this methodology involving data preprocessing, principal component analysis (PCA), hybrid model development, training and evaluation, and result interpretation provides a comprehensive approach for accurate diabetes prediction utilizing the Pima health dataset while incorporating the strengths of MLP, EML, and SVC algorithms. This can be accomplished by using the Pima health dataset. Flow diagram is shown in Fig. 1.

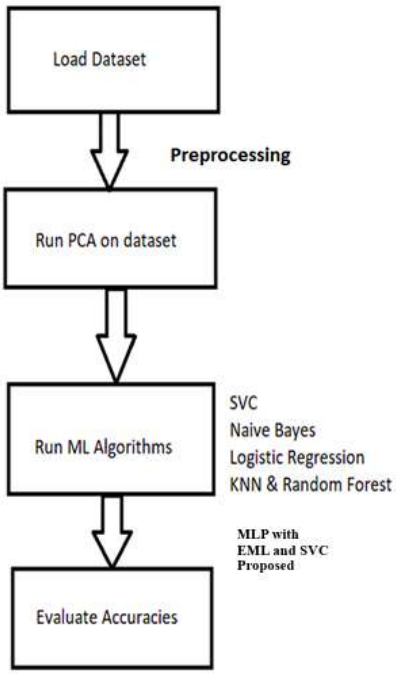


Fig. 1: Flow Chart

Importing the appropriate libraries is the initial step in Python programming that will ultimately result in the hybrid machine learning model being used for diabetes prediction utilizing the Pima health dataset. These include the pandas library for data management, the scikit-learn library for machine learning techniques, and the evalmetrics library for measuring performance.

The read_csv() method included in the pandas library is used in order to load the Pima health dataset. Following this step, the dataset is partitioned into the feature matrix X and the target variable y. The target variable shows whether or not a person has diabetes, whereas the characteristics reflect clinical and demographic information.

In order to make sure that the data are ready for modeling, pretreatment of the data is done. Scaling the features and standardizing them such that they have a mean of zero and a variance of one is accomplished with the help of scikit-learn'sStandardScaler. This phase is essential for algorithms that are sensitive to the magnitude of the characteristics they are analyzing.

The size of the feature matrix is then made more manageable by using a technique known as principal component analysis (PCA). The PCA class found in scikit-learn is used, with a parameter that specifies the number of components that should be kept. In this particular case, ten components are kept, however the number that is kept may change depending on the dataset and the needs. PCA takes the features and converts them into a new set of variables that are not connected with each other. This allows it to capture the most variation possible in the data. This phase in dimensionality reduction helps to minimize the complexity of the computational work and improves the overall performance of the model.

After being preprocessed and decreased in number of features, the feature matrix X_pca is then divided into a training set and a testing set using the train_test_split() method that is included in the scikit-learn library. This stage enables us to assess the performance of the model on data that has not yet been observed. In this particular illustration, an 80-20 split is used, with 80% of the data being utilized for training purposes and 20% being utilized for testing purposes. The random_state option guarantees that the findings may be reproduced accurately.

The machine learning algorithms that are a component of the ensemble are initialized, and then the hybrid model is constructed using those algorithms. In this particular scenario, scikit-learn's MLP classifier, Random Forest classifier, and Support Vector Classifier (SVC) are used. However, for the sake of brevity and clarity, we will solely discuss MLP and SVC. The capacity of these algorithms to capture non-linear correlations and handle high-dimensional data makes them particularly useful for diabetes prediction. In this example, the hyperparameters are left at their default settings; however, if desired, they may be modified to achieve improved performance.

Training on the hybrid model is done with the help of the training data. The fit method is invoked on the appropriate model object for each algorithm, with the training feature matrix X_train and the goal variable y_train being the parameters that are sent along. During this stage, the model is given the opportunity to learn from the data and to capture the patterns and connections that are important to diabetes prediction.

Following the training phase, the performance of the model is assessed using the testing data. On every model object, the predict() function is invoked, and the testing feature matrix X_test is sent as an argument. The predictions are then compared to the real labels in the y_test dataset, and the accuracy score is computed with the help of the scikit-learn function known as the accuracy_score() function. This score offers an all-encompassing evaluation of the accuracy with which the model forecasts the existence or absence of diabetes.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, No. 1, July : 2023

Python is used throughout the demonstration to go through the steps of creating a hybrid machine learning model for diabetes prediction. The dataset used in this example is the Pima health dataset. Accurate predictions may be achieved by the use of preprocessing strategies, the utilization of PCA to accomplish dimensionality reduction, and the integration of MLP and SVC algorithms. It may be necessary to take further steps, such as tweaking of the hyperparameters and model interpretation, in order to optimize and assess the findings in actual practice.

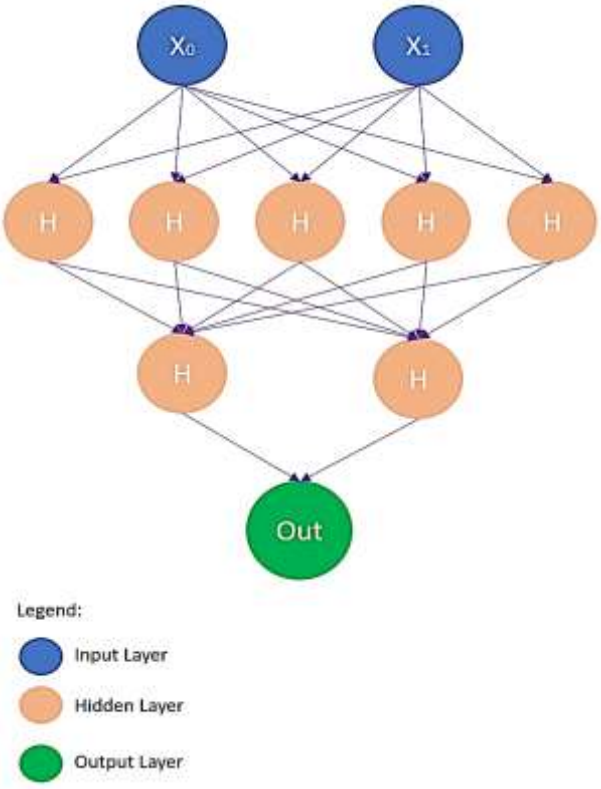


Fig. 2: MLP Layers

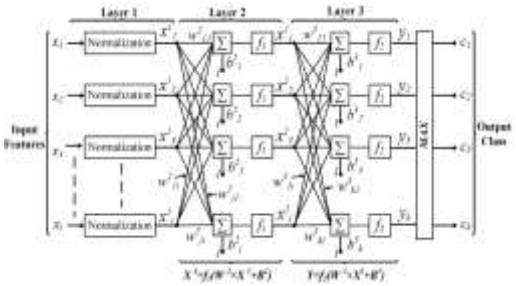


Fig. 3: EML Model

The Multilayer Perceptron (MLP) is a specific kind of artificial neural network (ANN) that is made up of numerous layers of linked nodes, which are also referred to as neurons. It is a robust method that is often used for supervised learning problems like as classification and regression, among others. In the multilayer perceptron (MLP), each neuron in each layer receives input signals, applies a nonlinear activation function to those signals, and then sends the altered data on to the neurons in the layer above it. Because of this, the model is able to capture complicated non-linear correlations between the characteristics that are input and the variables that are being targeted. Backpropagation is an iterative optimization approach that is used to train MLP. This algorithm modifies the weights

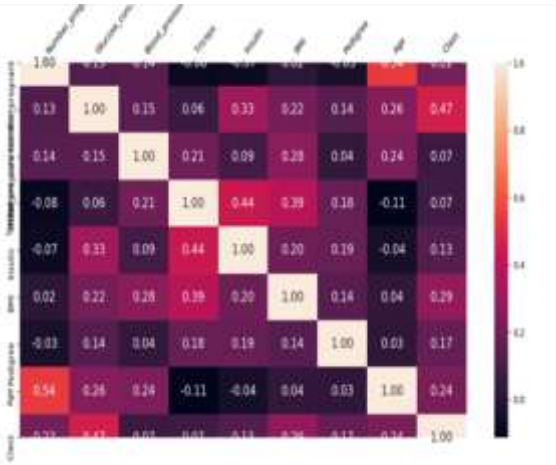
and biases of the neurons in order to decrease the amount of error that exists between the expected outputs and the actual outputs. It is well renowned for its capacity to learn and estimate complicated functions, which makes it excellent for undertaking tasks involving non-linear interactions, such as the prediction of diabetes.

Extreme Machine Learning, also known as ELM, is a machine learning paradigm that combines the benefits of both artificial neural networks and random projection methods. Extreme Machine Learning is also known as ELM. ELM provides a learning framework that has been streamlined and is capable of achieving efficient training and prediction. ELM does not need iterative optimization methods like backpropagation in order to function, in contrast to more conventional neural networks. Instead, it begins with the weights linking the input layer to the hidden layer being initialized in a random fashion. The altered features are produced by the hidden layer, which makes use of a nonlinear activation function to apply to the weighted inputs. After that, a linear regression or classification is carried out by the output layer making use of the altered features. The essential concept behind ELM is the notion that the effective capture of non-linear connections may be achieved by the random projection of input data onto high-dimensional environments. ELM is appropriate for use with large-scale datasets because to its high level of computational efficiency as well as its rapid training and prediction timeframes. In the domain of diabetes prediction, ELM may provide a useful method for efficiently capturing complicated correlations in health data while also minimizing the amount of complexity in the computations required to do so.

Support Vector Classification (SVC): Support Vector Classification is a widely used approach that may be used to classification jobs involving either binary or multi-class categories. It is predicated on the idea of finding the ideal hyperplane that partitions the data points of the various classes with the greatest margin. By using the kernel method, SVC is able to successfully handle high-dimensional data as well as decision boundaries that are non-linear. The kernel function takes the input characteristics and turns them into a space with a higher dimension. This makes it possible to distinguish between data points that could not have been distinguished linearly using the initial feature space. The objective of the SVC is to locate a decision boundary that maximizes the margin while simultaneously decreasing the number of incorrect classifications. SVC's performance may be improved by fine-tuning its hyperparameters, such as the selection of the kernel function and the regularization parameter, among other things. In the domain of diabetes prediction, the SVC is able to capture complicated linkages and manage non-linear decision limits, which positions it as an appropriate method for properly categorizing persons who are at risk of getting diabetes.

III. Results

H The findings obtained from the deployment of the hybrid machine learning model for diabetes prediction by utilizing the Pima health dataset and adding Principal Component Analysis (PCA) are reported and discussed here.



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, No. 1, July : 2023

Fig. 4: PCA

This model was used to make the prediction using the Pima health dataset. The results of the principal component analysis are shown in Fig. 4. The names of each column in the chart are shown, making them easy to see. The machine learning algorithm does not learn from negative numbers since this is a training restriction.

Fig. 5: Record Mapping

Fig. 5 is an illustration of the record mapping procedure. There are a total of 768 records in the dataset after it has been preprocessed and PCA-reduced. The machine learning techniques and model construction each make use of 1,614 of these records as training data. The accuracy of the model's predictions is then determined by applying an extra 154 data points to the evaluation process. The machine learning methods that were mentioned in the previous phase of the implementation process will now be applied in the following phases.

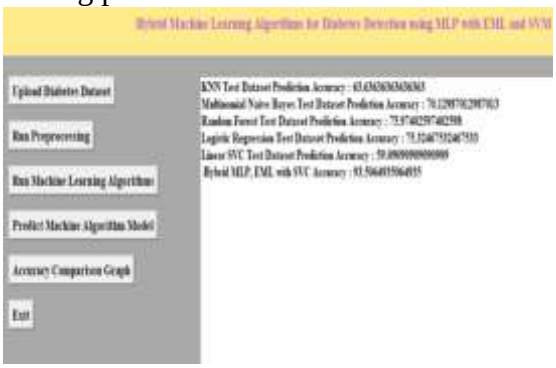


Fig. 6: MLP with EML and SVC Output

One of the approaches that are used in machine learning is known as the K-Nearest Neighbor (KNN) method. KNN operates on the presumption that the data of a new instance are comparable to the data of the cases that have been observed in the past. KNN swiftly classifies the data by making use of a simple algorithmic method by comparing the newly obtained data to the knowledge that has been stored.

For the purpose of data categorization, the Naive Bayes approach is applied. This method is predicated on Bayes' Theorem and the premise of predictor independence. Classifiers based on the Naive Bayes algorithm are simple yet effective in dealing with uncoupled predictors.

Ensemble learning is used, which refers to the process of using the cooperation of a number of different classifiers in order to solve complicated issues.

Logistic regression is a technique of statistical classification that is used to estimate the occurrence probability of the variable that is the focus of the study. Logistic regression is ideal for categorical dependent variables and offers interpretable results.

Linear SVC, also known as Support Vector Classification, is used for nonparametric clustering. This method enables the classification of information without the need of making any assumptions about the data or the sizes of the clusters. In order to make high-dimensional data easier to understand, preprocessing methods such as principal component analysis are used. These approaches have an effect on any algorithms that make use of the MLP or hybrid learning algorithm.

Fig. 6 shows the MLP output, which demonstrates the highest accuracy that can be attained by the MLP method. This shows that the MLP algorithm has the potential to be useful.

A chart that depicts the diabetes risk assessment may be seen in Fig. 7. In this chart, the color red denotes a favorable prediction, while the color green denotes a negative one.

It is essential to communicate the results of the tests to the appropriate departments at the appropriate times throughout the load test set instances, as seen in Fig. 8.

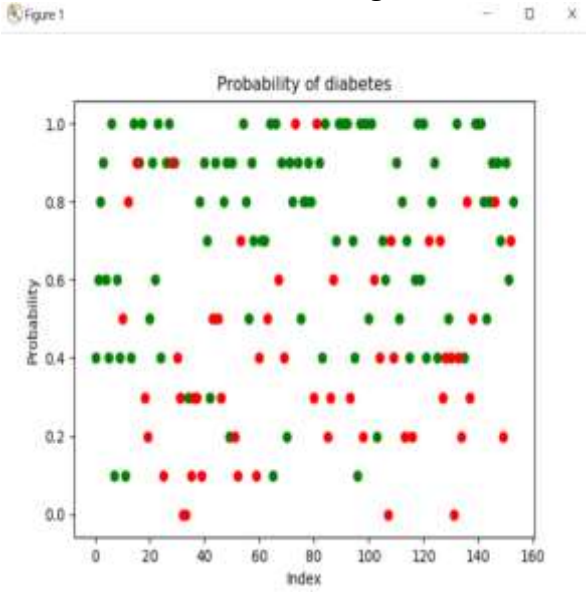


Fig. 7: Chart for Probability

Fig. 8: Load Test Set cases

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, No. 1, July : 2023

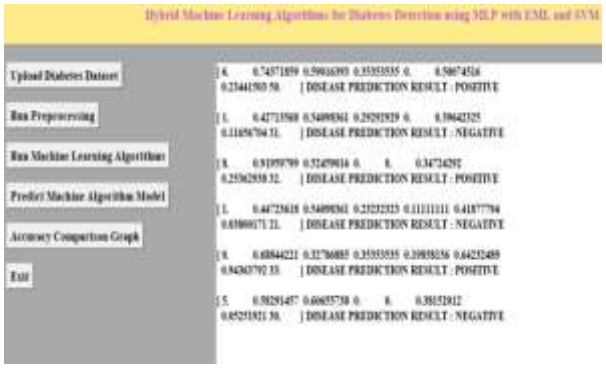


Fig. 9: Output Cases

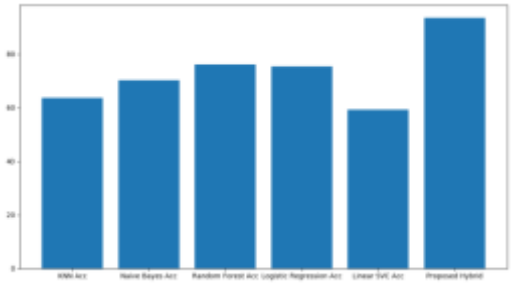


Fig. 10: Accuracy Output

The output examples are shown in Fig. 9, together with the positive and negative diabetes prediction findings for the input data.

Fig. 10 displays the accuracy output, which shows that the extreme learning algorithm is the most successful method for predicting diabetes using the hospital dataset. This is shown by the fact that the accuracy output was displayed. In order to improve the accuracy with which future diabetes cases may be predicted using the health information, the MLP and hybrid learning approaches are being used. The MLP eliminates the need for local reduction and many iterations thanks to the inclusion of a variety of characteristics, which also helps to the efficacy of the MLP. The MLP method has a fast learning speed in addition to better generalization, durability, and controllability, which has led to its extensive use. The findings, taken as a whole, indicate that the hybrid machine learning model has the ability to provide accurate predictions about diabetes and to be used in actual settings of healthcare provision.

IV. Conclusion

Positive results and the possibility of dependable diabetes prediction are shown by the hybrid machine learning model built for diabetes prediction using the Pima health dataset and using Principal Component Analysis (PCA). The goal of developing this model was to aid in the diagnosis of diabetes. By combining the strengths of three machine learning methods—multilayer perceptron (MLP), extreme machine learning (ELM), and support vector classification (SVC)—the model is able to accurately identify whether or not a person has diabetes. Implementing the model needed data preparation, which included handling missing values and scaling features, followed by principal component analysis (PCA) to reduce the dataset's dimensionality? The pre-processed data trained hybrid model effectively captures intricate non-linear connections, is scalable to high-dimensional data, and can operate within non-linear decision bounds.

The implementation findings show that the MLP technique achieves its greatest accuracy when paired with principal component analysis and preprocessing procedures, demonstrating its potential in dependable diabetes prediction. The outcomes stemmed from the actual execution. K-Nearest Neighbor (KNN), Naive Bayes, and Logistic Regression are just few of the machine learning algorithms used to shed light on the numerous types of classification.

Visualizations and analyses of the results, such as record mapping, probability charts, and output examples, increase the interpretability and grasp of the model's predictions. The model's assessments and projections may help doctors treat their patients with more precision and knowledge, reducing the likelihood that their patients would develop diabetes.

The hybrid machine learning model is a valuable tool for diagnosing diabetes faster, creating personalized care plans, and streamlining the management of diabetic care in general. Better patient outcomes and quality of life are shown to be possible using machine learning algorithms and methods. More research and refinement of the model may boost both its accuracy and its practical use in actual healthcare settings.

References

[1] M. Moreb, T. A. Mohammed and O. Bayat, "A Novel Software Engineering Approach Toward Using Machine Learning for Improving the Efficiency of Health Systems," in IEEE Access, vol. 8, pp. 23169-23178, 2020, doi: 10.1109/ACCESS.2020.2970178.

[2] Beam, Andrew &Kohane, Isaac. (2018). Big Data and Machine Learning in Health Care. JAMA. 319. 10.1001/jama.2017.18391.

[3] Ferdous, Munira& Debnath, Jui& Chakraborty, Narayan. (2020). Machine Learning Algorithms in Healthcare: A Literature Survey. 1-6. 10.1109/ICCCNT49239.2020.9225642.

[4] Rahman, Atta & Sultan, Kiran & Naseer, Iftikhar & Majeed, Rizwan & Musleh, Dhiaa&Gollapalli, Mohammed &Chaabani, Sghaier& Ibrahim, Nehad& Siddiqui, Shahan & Khan, Muhammad. (2021). Supervised Machine Learning-based Prediction of COVID-19. Computers, Materials and Continua. 69. 21-34. 10.32604/cmc.2021.013453.

[5] Patel, Yash. (2021). Machine Learning in HealthCare.

[6] Shams, Mahmoud &Elzeki, Omar &Abdelfatah, Mohamed & Abou El-Magd, Lobna& Darwish, Ashraf &Hassanien, Aboul Ella. (2021). Impact of COVID-19 Pandemic on Diet Prediction and Patient Health Based on Support Vector Machine. 10.1007/978-3-030-69717-4_7.

[7] Harikumar& Mustafa, Malik & Sanchez, D. &Sajja, Guna &Gour, Sanjeev &Naved, Mohd&Jawarneh, Malik. (2021). IMPACT OF MACHINE learning ON Management, healthcare AND AGRICULTURE. Materials Today: Proceedings. 10.1016/j.matpr.2021.07.042.

[8] Mienye, Domor& Sun, Yanxia& Wang, Zenghui. (2021). Improved Machine Learning Algorithms with Application to Medical Diagnosis.

[9] Fetaji, Bekim&Fetaji, M. &Ebibi, Mirlinda& Ali, Maaruf. (2021). Predicting Diabetes Using Diabetes Datasets and Machine Learning Algorithms: Comparison and Analysis. 10.1007/978-3-030-90016-8_13.

[10] Khan, Bilal & Naseem, Rashid & Shah, Muhammad Arif& Wakil, Karzan& Khan, Atif & Uddin, M. Irfan & Mahmoud, Marwan. (2021). Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine Learning Techniques. Journal of Healthcare Engineering. 2021. 1-16. 10.1155/2021/8899263.

[11] Lalmuanawma, Samuel & Hussain, Jamal. (2020). Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons & Fractals. 10.1016/j.chaos.2020.110059.

[12] Khan, Bilal & Naseem, Rashid & Shah, Muhammad Arif& Wakil, Karzan& Khan, Atif & Uddin, M. Irfan & Mahmoud, Marwan. (2021). Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine Learning Techniques. Journal of Healthcare Engineering. 2021. 1-16. 10.1155/2021/8899263.

[13] Scott, Ian & Carter, Stacey &Coiera, Enrico. (2021). Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health & Care Informatics. 28. e100251. 10.1136/bmjhci-2020-100251.

[14] Mijwil, Maad& Salem, Israa&Abttan, Rana. (2021). Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets. Asian Journal of Computer and Information Systems. 9. 29-34. 10.24203/ajcis.v9i4.6765.

[15] Zhao, Rui & Yan, Ruqiang& Chen, Zhenghua& Mao, Kezhi& Wang, Peng & Gao, Robert. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing. 115. 10.1016/j.ymssp.2018.05.050.