DEVELOPMENT OF A PREDICTIVE MACHINE LEARNING MODEL FOR DIABETES USING STACKED ENSEMBLE APPROACH

Thompson Ekemini[1], Joshua Agbogun PhD[2], Nwogwugwu Benjamin[3]
[1,2,3] Godfrey Okoye University, Enugu State, Nigeria

Article Info
Received: 11/04/ 2023
Revised: 27/05/2023
Accepted 20/06/2023

Corresponding Author's Email: ekemini.thompson@outlook.com
Corresponding Author's Tel: +2348085434165

ABSTRACT

Diabetes is a global health concern with millions of new cases annually. Early detection of the disease can prevent its progression and complications. In this study, we developed a prediction model that uses diagnostic measurements to determine if a patient has diabetes. To improve the model's performance and accuracy, we explored different techniques instead of relying on a single algorithm or dataset, which may not be optimal for the input data or parameters. We employed Logistic Regression and Stacked Ensemble Technique, and two feature selection methods, using two datasets: the PIMA Indians Diabetes dataset and a dataset from Enugu State University Teaching Hospital. Our results show that ensemble methods improve accuracy and prediction compared to a single model. The highest accuracy achieved was 79% for Dataset 1, while employing the stacked ensemble model on Dataset 2 resulted in a 99% accuracy in predicting the blood sugar disease. Our study demonstrates the benefits of using multiple algorithms and ensemble techniques to develop accurate diabetes prediction models.

Keywords: **Diabetes prediction, blood glucose level, Stacked Ensemble model, Enugu State University Teaching Hospital, Diabetes dataset, Logistic Regression**

1.0 INTRODUCTION

Diabetes is a pervasive disease that poses a significant threat to human health worldwide. This condition is characterized by elevated blood glucose levels resulting from defective insulin secretion, impaired biological effects of insulin, or a combination of both (Liu et al., 2018). The prevalence of diabetes has been on the rise globally, with the International Diabetes Federation projecting that the number of people living with this disease will double from the current estimate of 400 million to 800 million by 2035 (Pradhan et al., 2015). The increase in diabetes cases can be attributed to factors such as sub-urbanization, the adoption of unhealthy lifestyles, and an aging population that lacks adequate preparation for prevention and control, presenting significant challenges to diabetes care worldwide (Liu et al., 2018). According to the World Health Organization (WHO), 8.5% of adults aged 17 years and above are diabetic patients globally (WHO, 2020). In 2020 alone, diabetes was responsible for 1.5 million deaths, while high blood glucose led to 2.3 million deaths (Pradhan et al., 2015).

The prevalence of diabetes has doubled over the past decade, with over 400 million people affected globally and an annual incidence of 7% (Zahran, 2020; Sneha & Gangil, 2020). Early detection and prompt diagnosis of symptoms are essential to preventing severe consequences, but this remains a challenge in some cases (Temurtas et al., 2019). Recent studies have explored the use of machine learning models for predicting various diseases, with some clinicians currently using these models to predict different health conditions (Sisodia et al., 2019; Kandhasamy et al., 2015; Yuvaraj et al., 2017; Mercaldo et al., 2017). Developing a diabetes predictive model that is convenient, accurate, and cost-efficient is therefore crucial. Artificial intelligence techniques provide valuable insights for human-related fields such as medical diagnosis, which can be a challenging and time-consuming process (Mercaldo et al., 2021; Adeloye et al., 2018).

In recent years, numerous methods and algorithms have been developed for mining biomedical datasets for hidden information, including supervised learning techniques like Neural Networks (NNs), Fuzzy Logic Systems, Decision Trees (DT), Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and logistic regression. Unsupervised learning techniques like clustering analysis, pattern recognition, and image analysis, as well as reinforcement algorithms used in game theory, control theory, and decision theory, have also been explored (Modern, 2019; Mercaldo et al., 2017). This study proposes a prediction model for determining whether a patient has diabetes based on specific diagnostic measurements in the dataset. It explores various techniques to enhance the model's performance and accuracy using supervised learning algorithms and a dataset from the Enugu State University of Science and Technology Teaching Hospital. This research has contributed to the health sector by providing patients with accurate prior knowledge of their health status related to diabetes, reducing the incidence of complications, morbidity, and mortality associated with this disease.

2.0 BACKGROUND AND RELATED LITERATURES

Diabetes is fast becoming more and more common in people's daily lives with the prevalent cost and standard of living globally. Developing a predictive machine learning model for rapid and accurate diagnosis and analysis of diabetes is a topic worth researching. Medically, diabetes is diagnosed based on fasting blood glucose, glucose tolerance, and random blood glucose levels (American Diabetes Association, 2015). The sooner the diagnosis, the easier it is for us to manage the disease. Machine learning can help people make a preliminary assessment of diabetes from daily physical examination data and may be helpful to physicians (Lee and Kim, 2018). The most important issues in the machine learning process are the selection of valid properties, prediction model and the correct classifier. Several algorithms have been used to predict diabetes, including traditional machine learning methods such as support vector machines (SVMs), decision trees (DTs), and logistic regression (Kavakiotis et al., 2017).

Pradhan et al in 2015, trained and tested a database using genetic programming (GP) and predicted diabetes using diabetes dataset obtained from the UCI repository. According to Pradhan et al. (2015), genetic programming yields superior accuracy compared to other techniques that were used. Although it reduces the time it took to generate the classifier, it significantly did not improve accuracy. Kim, S. (2015) proposed a deep network structure using SVM with CPON to provide proper structural depth and robust classification accuracy in diabetes prediction. To simulate the proposed model, the Wisconsin breast cancer dataset, the Pima Indian diabetes dataset, the BUPA liver disease dataset, and the ionosphere dataset from the UCI machine learning repository and the MNIST dataset were tested.

The entire test datasets yielded the following accuracies: BUPA liver disorders, 77.14%; ionosphere, 97.22%; Wisconsin breast cancer, 98.55%; Pima Indians diabetes, 83.11%; and MNIST, 94.84%. The machine learning algorithm is well known in medicine for predicting disease. Many researchers use ML techniques to predict the best and most accurate results. Kandhasamy and Balamurali (2015) used several classifiers SVM, J48, K-nearest neighbor method (KNN), and random forest. The classification was performed using the data records in the UCI repository. Classifier results were compared based on accuracy, sensitivity, and specificity values. The classification was done in two cases, one when the dataset was preprocessed and the other when it was not preprocessed using 5-fold cross-validation. The author did not explain the pretreatment steps applied to the dataset, but just stated that the data had been denoised. They reported that the decision tree classifier J48 showed the highest accuracy of 73.82% without pretreatment, and the classifier KNN (k = 1) and Random Forest showed the highest accuracy of the 100th pretreatment of the data.

In 2015, Tafa introduced an enhanced, integrated model of Support Vector Machine (SVM) and Naïve Bayes for diabetes prediction. The model's performance was assessed using a dataset gathered from three distinct locations within Kosovo. The dataset consisted of 402 patients and eight attributes, with 80 patients being diagnosed with type 2 diabetes. Some attributes utilized in the study have not been investigated before, including the regular diet, physical activity, and family history of diabetes. The authors didn’t mention whether the data was preprocessed or not. The validation test divides the dataset into 50% for each training set and test set. The proposed combination algorithm improved the prediction accuracy to 97.6%. This score was compared to SVM and Naive Bayes performance, with scores of 95.52% and 94.52%, respectively. Mercaldo et al. (2017) used 6 different classifiers: J48, Multi-layer Perceptron, Hefting Tree, JRip, BayesNet, and Random Forest. The Pima Indian dataset was also used in this study. The author does not mention preprocessing steps, but uses two algorithms, Greedy Stepwise and Best First, to determine the identification attributes that help improve classification performance. Four attributes were selected: the classification of obesity index, plasma glucose levels, diabetic pedigree function, and age. 10x cross-validation is applied to the dataset. Comparisons between classifiers were based on precision, recall, and F-measure values. The results show that using the Hoeffding Tree algorithm, the accuracy value corresponds to 0.757, the recall corresponds to 0.762, and the measured value was 0.759.

In addition to the other studies, Nengi and Jaiswal (2016) aimed to apply the SVM to predict diabetes. The Pima Indians and Diabetes 130 American datasets were used as a combined dataset. The aim of this research was to confirm the dependability of the findings, as many researchers relied on a solitary dataset. With 102,538 samples as dataset, the study focused on 49 attributes, of which 38,115 were categorized as negative samples and 64,419 as positive . However, the authors did not provide a discussion of the attributes employed in the study. The dataset was pre-processed by replacing the missing values and out of range data by zero, the non-numerical values are changed to numerical values, and finally the data is normalized between 0 and 1. Before applying the SVM model, the authors used different methods to select features. The F-select script in the LIBSVM package selected four attributes, while the wrapper and ranker methods (from the Weka tool) selected nine and 20 attributes, respectively. In the validation process, the author used a 10-fold cross-validation technique. Combined datasets were used to increase the reliability of diabetes prediction with 72% accuracy.

Deepti and Dilip (2018) identified diabetes using the Decision Tree, SVM, and Naive Bayes classifiers. The purpose was to identify the classifier with the highest possible accuracy. This study

used a Pima Indian dataset. The dataset is partitioned using 10-fold cross-validation. The author did not discuss data preprocessing. Performance was assessed using accuracy, precision, recall, and F major measurements. The highest accuracy was achieved by Naive Bayes, which achieved 76.30%. Orabi et al. in 2015 developed a diabetes prediction system. Its main purpose was to predict diabetes that a candidate will suffer at a particular age. The proposed system was based on the concept of machine learning using decision trees. The results were satisfactory but the system developed a model that used only decision trees to predict the onset of diabetes at a particular age.
Yuvaraj and Sripreethaa (2017) announced a diabetes prediction application that uses three different machine learning algorithms, including Random Forest, Decision Tree, and Naive Bayes. After pre-treatment, the research used the Pima Indian Diabetes Dataset (PID). The author discussed the method of information extraction used for feature selection to extract related features. Only 8 of the 13 key attributes were used. In addition, the research divided the dataset into 70% for training and 30% for testing. The results show that the Random Forest algorithm has the highest accuracy rate of 94%. A new deep learning approach to detect type 2 diabetes was published by A. Mohebbi et al. in 2017. The author demonstrated that it is possible to detect type 2 diabetic patients using Continuous Glucose Monitoring signals. To address the challenges of using deep learning technology in today's healthcare environment, the authors focused on deep learning, natural language processing, reinforcement learning, and generalized methods of computer vision. Soltani and Jafarian (2016) predicted diabetes using a stochastic neural network (PNN). The algorithm was applied to a Pima Indian dataset. The author did not use pre-treatment techniques. The dataset is divided into 90% for the training set and 10% for the test set. The proposed method achieved 89.56% and 81.49% accuracy in training and test data, respectively. Using 90% for training and 10% test is not really recommended.

Rakshit et al. (2017) predicted diabetes from a Pima Indian dataset using a Two-Class neural network. The author pre-processed the dataset by normalizing all sample attribute values using the mean and standard deviation of each attribute for numerical stability. In addition, we used correlations to extract related features. However, the author does not mention these discriminating features. The dataset was split into a training set of 314 samples and a test set of 78 samples. The outcomes generated by this model have achieved a peak precision rate of 83.3%, surpassing the precision levels observed in earlier studies. Mamuda and Sathasivam (2017) applied three supervised learning algorithms, including Levenberg Marquardt (LM), Bayesian Regulation (BR), and Scaled Conjugate Gradient (SCG). In this study, the author used the Pima Indian dataset (768 samples and 8 attributes) to evaluate performance. In the validation study, authors used 10-fold cross-validation to split the data into training and testing. The authors reported that Levenberg Marquardt (LM) performed best with a validation set based on a root-mean squared error (MSE) of 0.00025091.

Mohebbi et al. (2017) used logistic regression as the basis for multi-layered neural perceptron networks and traditional neural networks (CNNs). The goal was to identify diabetics using a signal dataset from Continuous Blood Glucose Monitoring (CGM). This study utilized a dataset of 9 patients, where each patient's continuous glucose monitoring (CGM) data was collected for 10,800 days, resulting in a total of 97,200 simulated CGM days. However, the attributes utilized in this study were not explicitly mentioned. The dataset was divided into training, validation, and test sets based on the Leave-one-patient-out mutual validation technique. In fact, the author only selected 6 patients for training and validation, and 3 patients for testing. CNN achieved the highest accuracy at 77.5%. Pham et al. (2017) applied three different ML techniques to datasets manually collected from a regional hospital in Australia. The dataset consists of 12,000 samples (patients) and includes 55.5% males. Using several pretreatment techniques (not mentioned in their article), the sample was purified and reduced to 7191 patients. For validation, the dataset was split into 2/3 for training sets, 1/6 for validation, and 1/6 for testing. The methods were Long Short-Term Memory (LTSM), Markov, and PlainRNN. Precision values were used to compare the performance of the techniques. With LTSM, the highest accuracy value of 59.6% was achieved.

In addition, Balaj et al. (2018) predicted two types of diabetes using recurrent neural networks (RNNs). The author used a Pima-Indian dataset with 768 samples and eight attributes. Attributes are ranked according to their most important, as described in their study "Glucose, BMI, Age, Pregnancy, Diabetes Pedigree Function, Blood Pressure, Skin Thickness, and Insulin". To validate the study, we divided the dataset into 80% for training and 20% for testing. The prediction accuracy for type 1 diabetes was 78%, while the prediction accuracy for type 2 diabetes was 81%. Moreover, Balaji and colleagues (2018) employed Recurrent Neural Network (RNN) to forecast both type 1 and type 2 diabetes. The study employed the Pima Indian dataset, which comprised of 768 samples and eight attributes ranked in descending order of importance as follows: Glucose, BMI, Age, Pregnancies, Diabetes Pedigree Function, Blood Pressure, Skin Thickness, and Insulin. To validate the findings, the dataset was divided into 80% training and 20% testing. The study achieved a prediction accuracy of 78% for type 1 diabetes and 81% for type 2 diabetes.

Lekha and Suchetha (2019) used one-dimensional modified CNN to predict diabetes based on breath signals. The authors collected a dataset for breath signals composed of 11 healthy patients, nine diabetic patients of type 2, and five diabetic patients of type 1. The attributes used in this dataset were displayed. No pre-processing was performed on the dataset. For the validation process, the authors used Leave one out Cross Validation. The performance was evaluated based on the Receiver Operating Characteristics (ROC) curve which reached 0.96. Researchers aimed during their quest to further improve the prediction mechanism, they built combined models in order to boost the accuracy. The models can be a combination of machine learning classifiers or a classifier with Artificial Intelligence optimizer. These models revealed a high accuracy. Chawan (2019) conducted a research aimed at developing a system that can accurately predict early diabetes in patients by combining the results of various machine learning technologies. The study used two different supervised machine learning methods - SVM and logistic regression, to predict diabetes. It took into account the seven characteristics of the patient. They concluded that SVM performed better with 77% (79%) accuracy than logistic regression, which had 78% (78%) performance accuracy.

In a study conducted by Uloko et al. (2019), the prevalence of risk factors for diabetes in Nigeria was investigated using a random effects model and subgroup-specific DM prevalence to explain inter- and intra-study heterogeneity. The study found that the prevalence of diabetes in Nigeria is increasing in regions of the country due to urbanization, sedentary lifestyles, aging, and unhealthy diets. The researchers recommended the urgent need for a national approach to the care and prevention of diabetes. Kaur & Kumari (2020) used machine learning algorithms and a multi-factorial dimensionality reduction algorithm to detect diabetes. The experimental results showed that the SVM linear model worked well with an accuracy of 0.89, and the Boruta wrapper algorithm was useful for feature selection to improve accuracy. Olaniyi and Adnan (2019) used a multilayer feed-forward neural network with the back-propagation algorithm to predict diabetes using the Pima Indian Diabetes Database. The dataset was normalized before processing, and an accuracy of 82% was achieved. Miotto et al. (2020) proposed a framework called Deep-Patient that uses a database of electronic health records to predict various illnesses, including diabetes. The author recommended pre-processing the dataset using PCA to improve predictive performance, and the accuracy reached 0.91.

Aishwarya and Vaidehi (2018) used multiple machine learning algorithms to predict diabetes, and logistic regression gave an accuracy value of 96%. Tejas and Pramila (2016) used logistic regression and SVM algorithms to build a diabetes prediction model and found that SVM performed better with an accuracy of 79%. Adnan (2018) designed a diabetes prediction model using three different machine learning algorithms, and the Random Forest algorithm achieved the highest accuracy rate of 84%. Deepti and Dilip (2018) used Decision Tree, SVM, and Naive Bayes algorithms and obtained the highest accuracy of 76.30% with the Naive Bayes algorithm. Finally, Sneha & Gangil (2020) used the predictive analytics tool WEKA to develop a supervised ML approach for early detection of diabetes, and the decision tree and random forest algorithms performed the best with an accuracy of 98.20% and 98.00%, respectively. The authors suggested future work on the use of rarely or non-used attributes for diabetes prediction.

3.0 METHODOLOGY:

The system analysis and design technique adopted in this study is the Rapid application development (RAD) technique. Rapid application development highlights speed and agility. This rapid pace is spearheaded by RAD’s capability and stress on minimizing the planning stage efforts and maximizing and fastening the prototype development and later helping in faster project release times. Rapid application development methodology helps in creating a production-ready application at a faster pace, while new functionality continues to be released at later stages. RAD reduces the risks of the waterfall model, with shortened cycle time and improved productivity and fewer resources. This greatly reduces the cost of application development.

3.1 Model Design

The algorithm used in the prediction model is Logistic Regression and Random Forest; Other machine learning techniques such as Decision Tree, Gradient Boost, Support Vector Machines, K-nearest Neighbors and Stochastic Gradient Descent classifier are used in the ensemble methods to test the improvement in the original performance. We have designed architecture of the logistic prediction model, which is shown in Figure 3.1. It shows the flow of how the implementation will be carried out. In addition various methods are explored to improve the performance and execution time. Firstly, it starts with two feature selection methods - Univariate Feature Selection and then the cross validation.

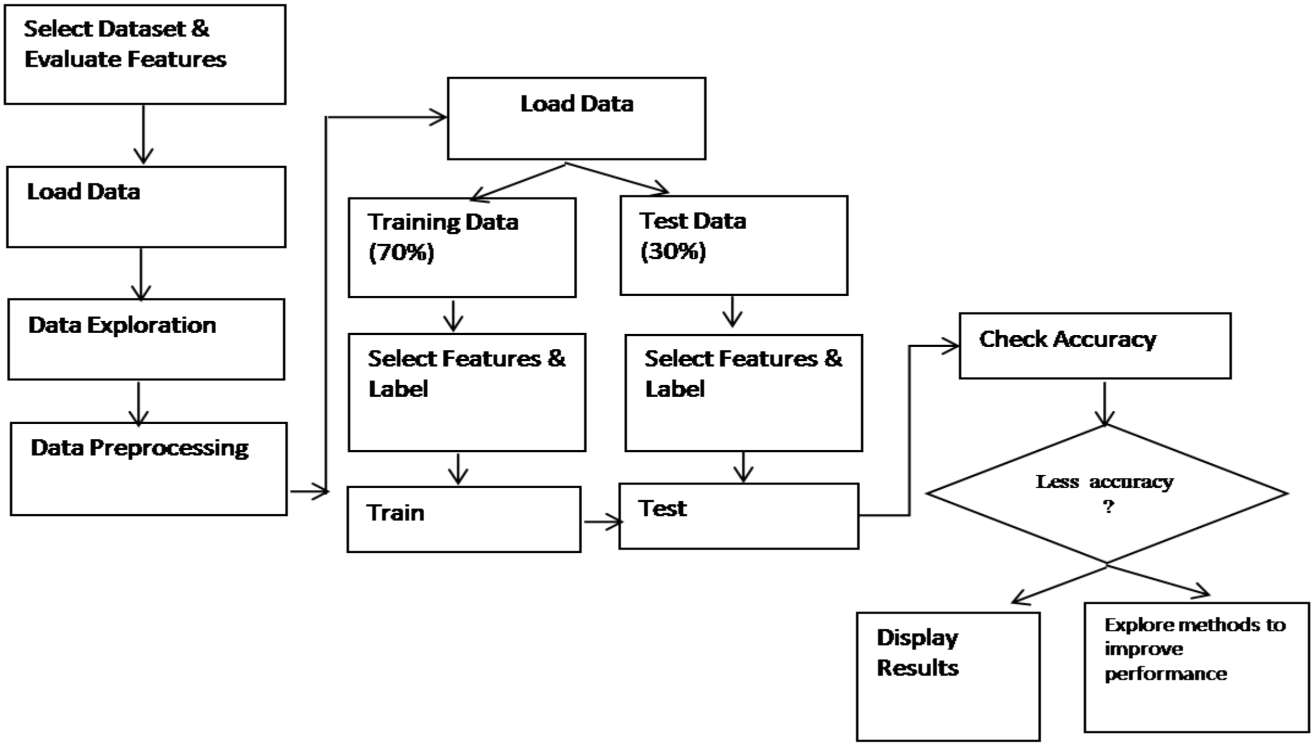

Figure 3.1: The architecture of the Logistic regression Diabetes Prediction Model.

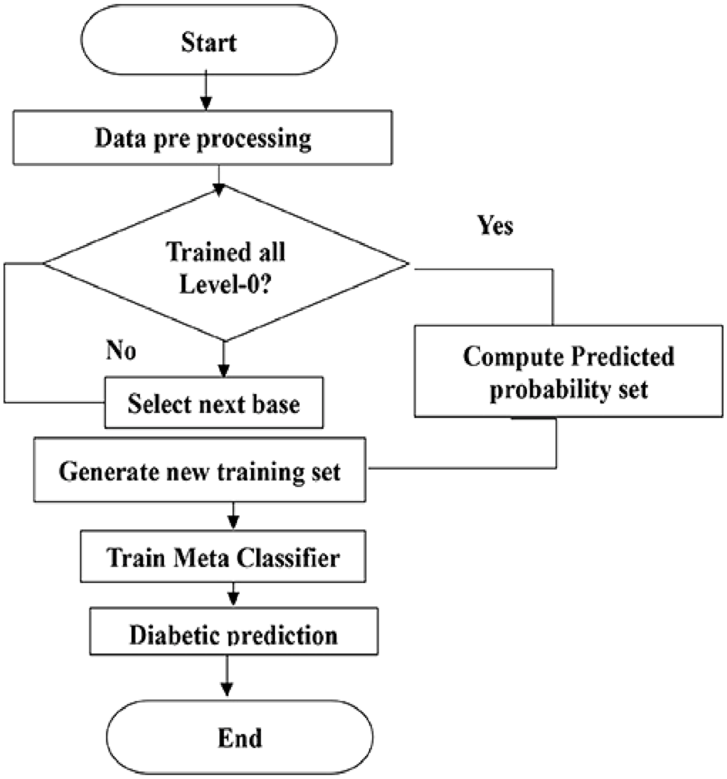

Figure 3.2 Proposed system flow chart.

This dataset consists of set predicted probabilities of each class of each classifier. A row r_i is the predicted probabilities base classifiers of each class of i_{th} row of the original dataset. The formula for the final prediction is done by using this Equation.

$$p = \frac{e^{bo+b1(x)}}{1 + e^{b0+b1\,(x)}}$$

$b0$, $b1$ are the constants, and x is the input vector. p is the final prediction, which is >0.5, then the patient has diabetic positive; otherwise, the patient has diabetic negative.

4.0 RESULT AND DISCUSSION:

Ensemble methods were further used to try and boost performance. Max/Majority Voting and Stacking methods were tested on both the datasets. The former method proved to be a best method among all, by showing a significant improvement in performance. The latter performed well after cross- validation was incorporated. Table 4.1, given below, shows the complete summary of the various techniques used in this experiment to build and improve the model, with their accuracy values.

Table 4.1: Comparison of Accuracy with cross-validation and Stacking

	Base Models	**Accuracy after k-fold**	**Final Model Accuracy of RF after K-fold**
Dataset 1	**LR** **Decision Tree** **Gradient Boost** **RF**	**0.77** **0.72** **0.77** **0.76**	**0.78**
Dataset 2	**LR** **Decision Tree** **Gradient Boost** **RF**	**0.91** **0.96** **0.97** **0.97**	**0.99**

In this ensemble experiment, we made use of seven different machine learning algorithms: Logistics Regression, KNN classifier, Decision tree, Gradient Boosting, RF, SVM, and SGD. The accuracy of each individual model, as well as the ensemble models for both datasets, is shown in Table 4.2.

Table 4.2 - Accuracy for individual and ensemble models.

	LR	**DT**	**SVM**	**KNN**	**RF**	**XGB**	**SGD**	**Ensemble**
Dataset 1	**0.77**	**0.75**	**0.75**	**0.75**	**0.81**	**0.78**	**0.70**	**0.79**
Dataset 2	**0.95**	**0.91**	**0.95**	**0.85**	**0.96**	**0.92**	**0.95**	**0.99**

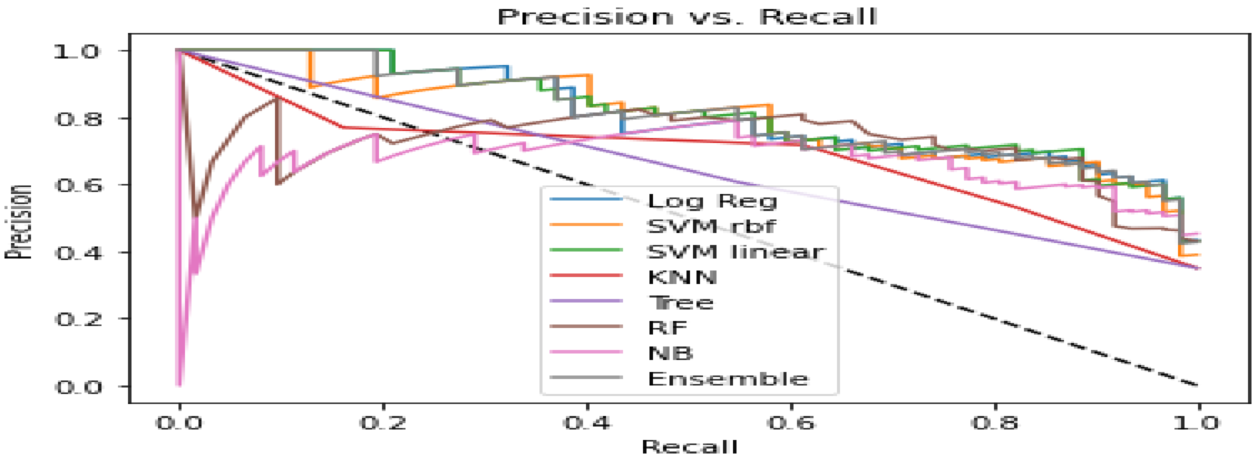

Figure 4.1. Precision-Recall Curve of the proposed system and various machine learning models.

A Precision-Recall Curve (PRC) is a metric used to compute the quality of the classifier model. The PRC curve is represented in a graph, where X-axis contains recall values Y-axis contains precision values. This curve depicts the compromise between precision and recall. The precision-recall curve shows the trade-off between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. In the graph of figure 4.1, high scores for both show that the ensembled model is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). A system with high recalls but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. In the system developed above, it shows a high precision and high recall which will return many results, with all results labeled correctly. The curve values are represented as TP/ (TP+FN) on the Y-axis.

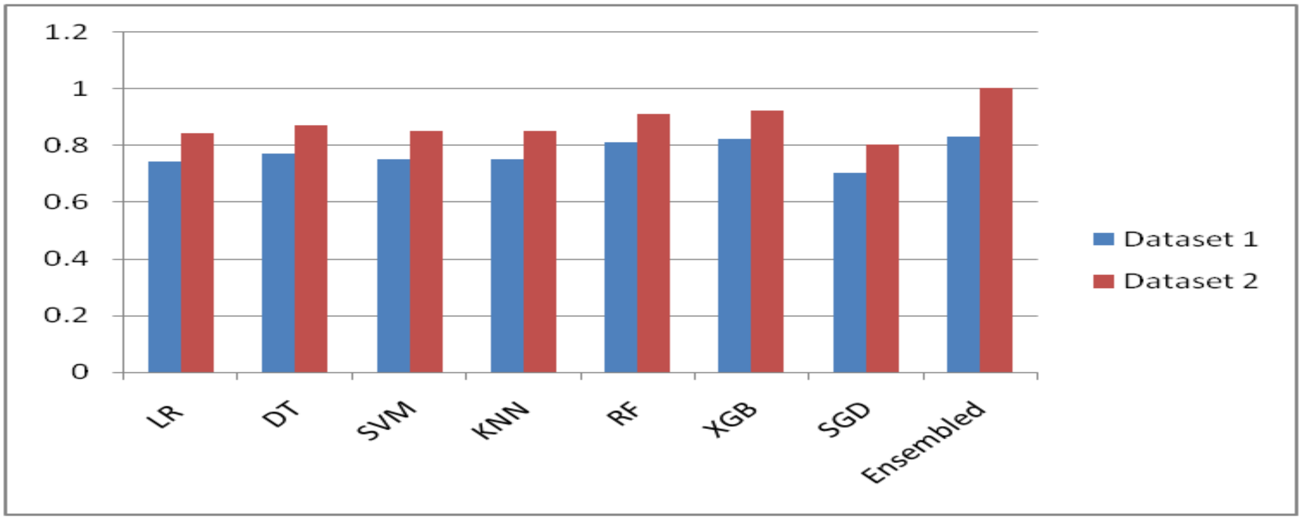

Figure 4.2. Accuracy chart of various models and proposed model comparison chart.

Data visualization and interpretation are very important to understand the data and its property. Making decisions from raw data is really difficult especially in machine learning, deep learning, accuracy comparison, etc. Using matplotlib library in python, the figure above shows Accuracy chart of various models and proposed model comparison chart.

Table 4.3 Quality metrics results.

Methods	**Precision**	**Recall**	**F1-Score**	**Accuracy**
RF	**0.74**	**0.78**	**0.69**	**0.96**
KNN	**0.84**	**0.71**	**0.85**	**0.85**
SGD	**0.62**	**0.74**	**0.68**	**0.92**
DT	**0.61**	**0.62**	**0.62**	**0.91**
LR	**0.61**	**0.74**	**0.67**	**0.95**
SVM	**0.59**	**0.74**	**0.66**	**0.77**
Stacking	**0.95**	**0.95**	**0.95**	**0.99**

The proposed system is compared with other machine learning models by quality metrics such as precision, recall, accuracy, and F1-score. These values are plotted in 12. The proposed stacked ensemble model obtained higher results compare to all other methods. Table 4.3 shows the quality metrics results in dataset 2. The proposed method is combination of machine learning algorithms. Generally multiple algorithms for a single problem shows better performance. Each machine learning model has its own strengths and weaknesses. If more than one model is combined, then the weakness may be averaged and strength will be increased for many problems, but not all problems. Thus the ensemble techniques such as bagging, boosting, and stacking are popular. Processing time can be higher than single algorithms. The proposed work is also tested with six machine learning approaches with different combination in ensemble technique and obtained lesser than 93% of accuracy of proposed approach.

Table 4.8: Comparison of Accuracy with and without Stacking

	Without Stacking	With stacking
Dataset 1	0.77	0.79
Dataset 2	0.95	0.99

The proposed stacked ensemble model obtained higher results compare to all other methods. Table 4.8 shows the Comparison of Accuracy with and without Stacking. The proposed method is combination of machine learning algorithms.

5.0 SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Summary

The literature has explored the use of AI techniques to predict diabetes, and this study builds upon prior work by applying several methods, including feature selection and cross-validation, to two datasets to improve accuracy. Unlike previous research that focused on a single model, this study used ensemble techniques with multiple models to improve performance. It was found that the algorithm used is not the only factor influencing performance, and that data preprocessing, feature selection, and ensemble techniques also play a role. However, challenges remain, including the need for larger datasets, reproducibility and external validation, and the risk of sensitive information being leaked. The proposed stacking ensemble model outperformed existing models in detecting diabetic positive patients, achieving 99% accuracy on a highly categorical dataset. The study recommends using Random Forest algorithms for classification and prediction problems, while also considering other algorithms that have competitive accuracy. These algorithms can be combined with other deep or machine learning techniques and AI to further improve accuracy and performance. The early detection of diabetes is crucial, and the proposed system offers a promising approach for reducing medical expenditure, death rates, and patient risk.

5.2 CONCLUSION AND FURTHER WORK:

Logistic Regression and Random Forest are established algorithms in the field of prediction modeling, known for their efficiency. However, the accuracy of a model depends on several factors, including data pre-processing. Effective pre-processing techniques involve removal of redundant and null values, as well as normalization of features with large scale differences.

In this study, we have identified feature selection and cross-validation as significant contributors to improving accuracy and reducing runtime. Ensemble techniques that combine various algorithms have also been found to enhance model performance. Cross-validation is especially important for boosting accuracy. The global challenge of diabetes affects individuals of all ages, and early detection is crucial for reducing medical expenses, mortality rates, and patient risk. The proposed system focuses on predicting the likelihood of diabetes early on, using the Pima Indians Diabetes Database (PIDD) and a dataset from ESUT Hospital. Our stacked ensemble model achieved 99% accuracy on a highly categorical dataset. Researchers are continually exploring new classifiers and models to enhance diabetes prediction accuracy. Early disease detection remains a vital approach in the medical field, given the increasing rates of diabetes patients worldwide and the lack of a vaccine for prevention. Future studies will aim to improve the model's ability to predict all possible complications, using an ordered sequence based on the likelihood of occurrence. Additionally, this work can be extended by incorporating other deep learning algorithms and techniques to automate diabetes type analysis.

6.0 REFERENCES:

Adeloye, D., Ige, J. O., Aderemi, A. V, Adeleye, N., Amoo, E. O., Auta, A., & Oni, G. (2018). Estimating the prevalence , hospitalisation and mortality from type 2 diabetes mellitus in Nigeria :a systematic review and meta-analysis. 1–16.

Adnan, K. (2018) Onset diabetes diagnosis using artificial neural network. Int. J. Sci. Eng. Res., 5, 754–759.

Aishwarya,M., and Vaidehi, V. (2018) "Diabetes prediction using machine learning algorithms."Procedia Computer Science 165: 292-299.[DOI:10.1016/j.procs.2020.01.047]

American Diabetes Association. 2015. Retrieved from http://www.diabetes.org/diabetes-basics/type-2 Chicago: "Type 2." . http://www.diabetes.org/diabetes-basics/type-2.

Balaji, H.; Iyengar, N.; Caytiles, R.D. (2018) Optimal Predictive analytics of Pima Diabetics using Deep Learning. Int. J. Database Theory Appl., 10, 47–62.

Chawan, P. M. (2019). Logistic Regression and Svm Based Diabetes. International Journal For Technological Research In Engineering, 5(6), 4347–4350.

Deepti S., Dilip S., (2018) Prediction of Diabetes using Classification Algorithms, Procedia Computer Science,Volume 132,2018,Pages 1578-1585,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.122.

Kaur, H., R. Kumari, Z. Ahmed (2020) Role of data mining in establishing strategic policies for the efficient management of healthcare system–a case study from Washington DC area using retrospective discharge data, BMC Health Services Res. 12 (S1) (2012) P12.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104–116. doi: 10.1016/j.csbj.2016.12.005

Kandhasamy, J.P.; Balamurali, S. (2015) Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Comput. Sci., 47, 45–51.

Lee, B. J., and Kim, J. Y. (2018). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. 20, 39–46. doi: 10.1109/JBHI.2015.2396520

Lekha, S.; Suchetha, M. (2019) Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network. IEEE J. Biomed. Health Inform, 22, 1630–1636.

Liu, J., Tang, Z. H., Zeng, F., Li, Z., & Zhou, L. (2018). Artificial neural network models for prediction of cardiovascular autonomic dysfunction in general Chinese population. BMC Medical Informatics and Decision Making, 13(1).

Mamuda, M.; Sathasivam, S. (2017) Predicting the survival of diabetes using neural network. In Proceedings of the AIP Conference Proceedings, Bydgoszcz, Poland, 9–11 May; Volume 1870, pp. 40–46.

Mercaldo, F., Nardone, V., Santone, A., 2017. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science 112, 2519–2528.

Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. (2020) Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci. Rep. 2016, 6, 26094.

Modern, S. (2019). A critical review on machine learning algorithms and their applications in pure sciences. Research Journal of Recent Sciences, 8(1), 14–29.

Mohebbi A., R., Chamkha, A., and Pop, I. (2017). Effects of Cavity and Heat Source Aspect Ratios on Diabetes and Natural Convection of a Nanofluid in a C-Shaped Cavity Using Lattice Boltzmann Method. Hff 28, 1930–1955. doi:10.1108/HFF- 03-2018-0110

Nengi O. and Jaiswal, F (2016) Choudhury, A.; Gupta, D. A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. In Recent Developments in Machine Learning and Data Analytics; Springer: Singapore; pp. 67–68.

Olaniyi, E.O.; Adnan, K. (2019) Onset diabetes diagnosis using artificial neural network. Int. J. Sci. Eng. Res., 5, 754–759.

Orabi, K.M., Kamal, Y.M., Rabah, T.M., (2015). Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420-427.

Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. (2017) Predicting healthcare trajectories from medical records: A deep learning approach. J. Biomed. Inform. 2017, 69, 218–229.

Pradhan, N., Rani, G., Dhaka, V. S., & Poonia, R. C. (2015). Diabetes prediction using artificial neural network. Deep Learning Techniques for Biomedical and Health Informatics,121, 327–339

Rakshit, S. Suvojit, M.; Sanket, B.; Riyanka, K.; Priti, G.; Sayantan, M.; Subhas, B. Prediction of Diabetes Type-II Using a Two-Class Neural Network. In Proceedings of the 2017 International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 March 2017; pp. 65–71.

Sisodia, D., Archenaa and Dr. E.A. Mary Anita, (2019) "Health Recommender System using Big data analytics", Journal of Management Science and Business Intelligence

Soltani, Z., & Jafarian, A. (2016). A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II. International Journal of Advanced Computer Science and Applications, 7(6), 89–94. doi:10.14569/IJACSA.2016.070611

Singh AK, Singh A, Shaikh A, Singh R, Misra A. Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India and other developing countries. Diabetes Metab Syndr. 2020;14(3):241-6.

Sneha, N., & Gangil, T. (2020). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data,vol. 6, no. 1.

Tafa, Zilbert. (2015) "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus," Bonfring International Journal on Data Mining, vol. 4, no.2, pp. 7-11.

Tejas H., and Pramila M., (2016) Intrusion detection system based on K-star classifier and feature set reduction. IOSR J. Comput. Eng., 15, 107–112.

Temurtas, H., Yumusak, N., & Temurtas, F. (2019). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4), 8610–8615.

Uloko AE, Musa BM, Ramalan MA, Gezawa ID, Puepet FH, (2019). Prevalence and risk factors for diabetes mellitus in Nigeria: a systematic review and meta-analysis. Diabetes Ther. ;9(3):1307–16.

World Health Organization (2020) Data and statistics http://apps/who.int/research/en/

Yuvaraj, N. & K.R. SriPreethaa, (2017) "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster", Springer.

Zahran, B. (2020). A Neural Network Model for Predicting Insulin Dosage for Diabetic Patients. International Journal of Computer Science and Information Security (IJCSIS), 14(6), 770–777