Extreme Gradient Boosting and Soft Voting Ensemble Classifier for Diabetes Prediction

Dayal Kumar Behera
Department of Computer Science & Engineering
Silicon Institute of Technology
Bhubaneswar, Odisha, India
dayalbehera@gmail.com

Shreela Dash
Department of Computer Science & Engineering
Centurion University of Technology and Management
Jatani, Odisha, India
shreelamamadash@gmail.com

Ajit Kumar Behera
Department of Computer Science & Engineering
Silicon Institute of Technology
Bhubaneswar, Odisha, India
ajitcs@silicon.ac.in

CH. Sanjeev Kumar Dash
Department of Computer Science & Engineering
Silicon Institute of Technology
Bhubaneswar, Odisha, India
sanjeevc@silicon.ac.in

2021 19th OITS International Conference on Information Technology (OCIT) | 978-1-6654-1664-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/OCIT53463.2021.00046

Abstract—**Diabetes is a chronic disease that has been impacting an increasing number of people throughout the years. Each year, it results in a huge number of deaths. Due to the fact that late diagnosis results in severe health complications and a significant number of deaths each year, it is critical to develop methods for early detection of this pathology. As a result, early detection is critical. Machine learning techniques aid in the early detection and prediction of diabetes. However, machine learning models do not perform well with missing values in the dataset. Imputation of missing values improves the outcome. In this work, we have used extreme gradient boosting (XGB) and another six traditional classifiers to train the model with a strong emphasis on missing value imputation. Results of various classifiers are combined using the soft voting ensemble (SVE) approach by assigning equal weights to the classifier. The experimentation uses the Pima Indian Diabetes Dataset, which contains information about people with and without diabetes. Missing value imputation for data pre-processing and ensemble classification has been shown to beat traditional classifiers.**

Keywords-Ensemble Learning, Missing value imputation, Diabetes Prediction, Soft Voting Classifier, XGBoost

I. INTRODUCTION

According to the “WHO (World Health Organization)”, for about 1.6 million people die each year from diabetes [1]. The disease Diabetes arises when the human body’s blood glucose/ sugar level is abnormally high. Type-1 diabetes, generally known as insulin dependent diabetes, is most frequently diagnosed in childhood. In Type-1, the pancreas is attacked by the body’s antibodies, which then kills internal body parts and causes the pancreas to stop producing insulin. Type-2 diabetes [2][3] is often referred to as adult-onset diabetes or non-insulin dependent diabetes. Although it is more merciful than Type 1, it is nevertheless extremely damaging and can result in serious complications, particularly in the small blood vessels of the eyes, kidneys and nerves [4]. Type-3 Gestational Diabetes [5] develops from an increase in glucose levels in women whose diabetes is not recognized earlier during pregnancy. The authors in [6] have created and validated a risk score for primary cesarean delivery in women with gestational diabetes. In women with GDM (gestational diabetes mellitus), a risk score based on nulliparity, excessive gestational weight gain, and usage of insulin can be used to determine the likelihood of primary cesarean delivery. M. Ghaderi *et al.* [2] worked on the effect of smartphone education on the risk perception of type 2 diabetes in a woman with GDM. Diabetes mellitus is related to long-term consequences. Additionally, people with diabetes face an increased chance of developing a variety of health concerns. Sugar levels in the human body typically range between 70 and 99 mg/deciliter [1]. If the glucose level is more significant than 126 mg/deciliter then the person is considered as a diabetic patient. Prediabetes is defined as a blood sugar level of 100- 125 mg/deciliter [7]. This disease is influenced by several factors such as height, weight, hereditary, and insulin [8], but the primary factor evaluated is glucose content. Early detection is the only approach to avoid difficulties. Predictive analytics strives to improve disease diagnosis accuracy, patient care, resource optimization, and clinical outcomes. Numerous researchers are conducting experiments to diagnose disease using variants of classification algorithms. Researchers have demonstrated that ML techniques [9][10] perform better at diagnosing various diseases. Bayes algorithm, Support Vector Machine, and DT (Decision Tree) machine learning classification algorithms are applied and assessed in work [8] to predict diabetes in a patient using the PIMA dataset. Machine learning techniques can also be utilized to identify individuals at elevated risk of Type-2 diabetes [11] or prediabetes in the absence of established impaired glucose regulation. BMI, waist-hip ratio, age, blood pressure (BP), and diabetes inheritance were the most impactful factors. Increased risk of Type-2 diabetes was associated with high levels of these characteristics and diabetes heredity. Machine learning techniques aid in the early detection and prediction

978-1-6654-1664-1/21/$31.00 ©2021 IEEE
DOI 10.1109/OCIT53463.2021.00046

of diabetes. However, ML models do not perform well with missing values in the dataset. This work emphasizes the missing value imputation. The objectives of this work are as follows:

- Study the impact of missing value in the PIMA diabetes dataset.
- Performing missing value imputation by replacing the missing value with the mean value of the group.
- Designing an extreme gradient boosting model for classifying diabetes.
- Comparative analysis of various traditional classifiers with the ensemble classifier.

Section II covers the related work. Information about the dataset and proposed model is presented in section III. Result analysis is done in section IV, and section V shows the conclusion of the work.

II. Related Works

The purpose of the article [12] is to illustrate the construction and validation of 10-year risk prediction models for Type-2 diabetes mellitus (T2DM). Data collected in 12 European nations (SHARE) are used for validation of the model. The dataset included 53 variables encompassing behavioural, physical, and mental health aspects of participants having an age more than fifty. To account for highly imbalanced outcome variables, logistic regression model was developed, each instance was weighted according to the inverse percentage of the result label. The authors used a pooled sample of 16,363 people to develop and evaluate a global regularized logistic regression model with AUC of 70 percent. Continuous Glucose Monitoring (CGM) devices continue to have a temporal delay, which can result in clinically significant differences between the CGM and the actual blood glucose level, particularly during rapid changes. In [13], authors have used the artificial neural network regression (NN) technique to forecast CGM results. Diabetes can also be a risk factor for developing other diseases such as heart attack, renal impairment, and partial blindness.

Kayal Vizhi and Aman Dash [14] worked on smart sensors and ML techniques such as RF and XGB (extreme gradient boosting) for predicting whether a person would get diabetes or not. A Mujumdar *et al*. [5] developed a diabetes prediction model that took into account external risk variables for diabetes including factors such as glucose, BMI, age, and insulin. The new dataset improves classification accuracy when compared to the available PIMA dataset. The study [15] covers algorithms such as linear regression, decision trees, random forests, and their advantages for early identification and treatment of disease. The research study discussed the predictive accuracy of the algorithms mentioned above.

Mitushi Soni and Sunita Varma[16] forecasted diabetes using Machine Learning Classification and ensemble approaches. When compared to other models, each model's accuracy varies. Their findings indicate that Random Forest outperformed different machine learning algorithms in terms of accuracy. Jobeda Jamal Khanam and Simon Y. Foo [1] conducted research using the PIMA dataset. The collection comprises data on 768 patients and their nine unique characteristics. On the dataset, seven machine learning algorithms were applied to predict diabetes. They concluded that a model combining LR and SVM is effective at predicting diabetes.

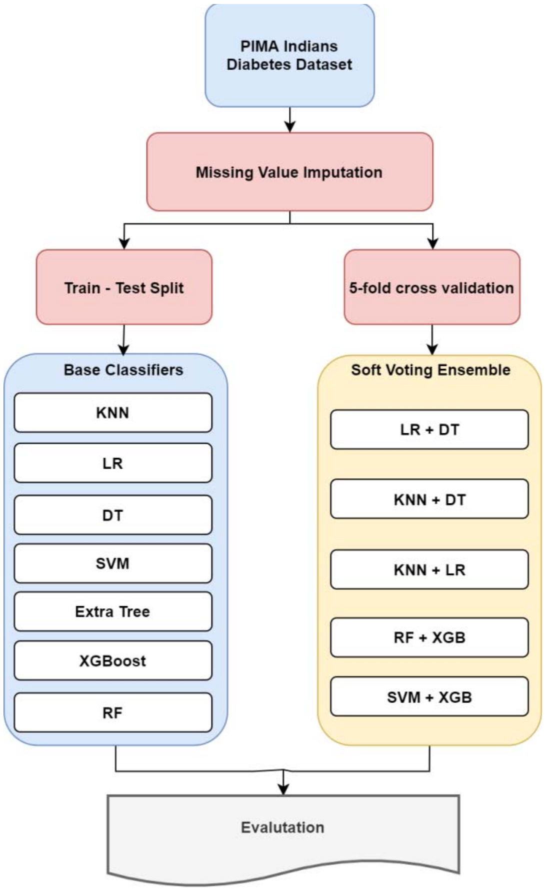


Figure 1. Proposed Framework for Diabetes Prediction.

Tibor V. Varga [17] used NCBI PubMed to conduct a systematic search. Articles that had the words "diabetes" and "prediction" were chosen. To illustrate the distinction between association and prediction, simulated data were constructed. It has been indicated that biomarkers with more effect sizes and small P-values might nonetheless have low discriminative utility. The article [18] attempts to synthesize the majority of the work on traditional ML and data mining techniques to predict diabetes and its complications. Hyperglycemia is a symptom of diabetes caused by insufficient insulin secretion and/or use. For experimental purposes, K. Kalagotla *et al*. [19] designed a novel stacking method based on multi-layer perceptron, SVM, and LR. The stacking strategy combined the intelligent models and improved model performance. In comparison to AdaBoost, the proposed unique stacking strategy outperformed other models. Authors in [20] worked on a pipeline for predicting diabetes individuals using deep learning techniques. It incorporates data enhancement using a variational autoencoder (VAE), feature enhancement via a sparse autoencoder (SAE), and classification via a convolutional neural network (CNN).

III. Materials and Methods

A. Dataset

In this work, the Pima Indians Diabetes dataset available in UCI ML repository is used for model evaluation. In the dataset all the patients are female and more than 21 years old. It contains a variety of medical predictor variables and one outcome variable. The pregnancies, BMI, insulin-level, age, BP, skin-thickness, Sugar/glucose, pedigree function are all independent variables. Outcome is the dependent variable.

B. Proposed Model

This research focuses heavily on enhancing the outcomes and accuracy of diabetes detection. The proposed approach is depicted in Fig 1. Numerous classical machine learning classifiers and related ensemble variation models are used to categorize disease as positive or negative. Numerous characteristics in the original dataset have entries of 0. According to the experts' advice, these values must be considered a missing value, such as sugar level, BP, skin thickness, insulin-level, BMI, pedigree function, and age can't be zero. Hence, the missing value is imputed by considering the mean value of the group. After that, the class label of the dataset is evaluated by taking two different values into account: Diabetic is set to 1 and non-diabetic by 0. The dataset is split into train set (70%) and test set (30%). The Pima dataset contains 768 samples. As a result of the train-test split, the number of examples in the train and test set is 537 & 231, respectively. The validation data is used to train the base classifiers in two scenarios. In the first case, the classifier is trained with missing value and in another case by missing value imputation.

Base classifiers chosen to train the model are KNN, LR, SVM, Decision Tree (DT), Random Forest (RF), Extra Tree and XGBoost (XGB) Classifier. 5-fold cross-validation is used in soft voting ensemble classifier (SVC). Base classifiers chosen for the soft voting ensemble have been depicted in Fig 1. 0.5 weight is selected for both the models in the SVC.

IV. Results and Discussion

The classification performance in the form of precision, recall, overall accuracy, F1-score and AUC is evaluated using the most traditional machine learning classifiers such as KNN, SVM, DT, LR, RF, XGBoost and Ensemble Models. All the models are implemented in python language by taking the default parameter setting of the Scikit-learn library.

Table I depicts the performance of various models on the validation data without considering missing value imputation (MVI), whereas Table II represents performance by considering missing value imputation. The data shows that the AUC of all the models improves a lot in missing value imputation.

TABLE I. F1-Score and AUC Without MVI

Model	F1-Score	AUC
KNN	0.61	0.71
LR	0.64	0.73
DT	0.57	0.67
SVM	0.63	0.69
XGBOOST	0.64	0.73
Extra Tree	0.63	0.72
RF	0.47	0.65
SVM+XGB	0.67	0.75
RF+XGB	0.65	0.73
KNN+LR	0.60	0.70
KNN+DT	0.57	0.68
LR+DT	0.57	0.68

TABLE II. F1-Score and AUC With MVI

Model	F1-Score	AUC
KNN	0.76	0.80
LR	0.67	0.74
DT	0.81	0.84
SVM	0.67	0.71
XGBOOST	**0.86**	**0.89**
Extra Tree	0.75	0.80
RF	0.78	0.82
SVM+XGB	0.85	0.88
RF+XGB	**0.85**	**0.87**
KNN+LR	0.79	0.83
KNN+DT	0.84	0.86
LR+DT	0.82	0.85

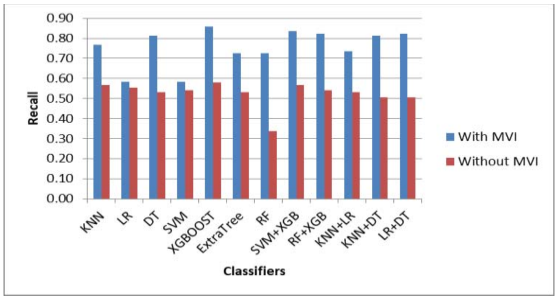


Figure 2. Recall Score of various classifiers.

The accuracy of ML classifiers can be derived from the confusion matrix. Fig.2 represents recall value of all the models with and without missing value imputation (MVI). Similarly, Fig. 3 depicts the precision score. The confusion matrix of one base classifier XGBoost and one Soft Voting Ensemble Classifier which combines the results of XGBoost and Random Forest is shown in Fig 4.

TABLE III. ACCURACY ON TRAIN-TEST SPLIT & K-FOLD CROSS-VALIDATION WITH AND WITHOUT MVI

Classifier	With MVI			Without MVI		
	Splitting Accuracy	K-Fold Accuracy-Mean	K-Fold Accuracy-Std	Splitting Accuracy	K-Fold Accuracy-Mean	K-Fold Accuracy-Std
KNN	0.81	0.84	0.02	0.74	0.72	0.02
LR	0.77	0.78	0.02	0.77	0.77	0.02
DT	0.85	0.89	0.02	0.71	0.67	0.03
SVM	0.77	0.78	0.02	0.77	0.77	0.02
XGBOOST	0.89	0.90	0.02	0.77	0.77	0.04
Extra Tree	0.81	0.86	0.03	0.77	0.77	0.03
RF	0.84	0.87	0.02	0.73	0.73	0.02
SVM + XGB	0.89	0.89	0.03	0.80	0.77	0.03
RF + XGB	0.88	0.90	0.02	0.79	0.77	0.03
KNN + LR	0.85	0.85	0.03	0.75	0.75	0.02
KNN + DT	0.87	0.88	0.03	0.72	0.70	0.04
LR + DT	0.85	0.88	0.02	0.72	0.68	0.02

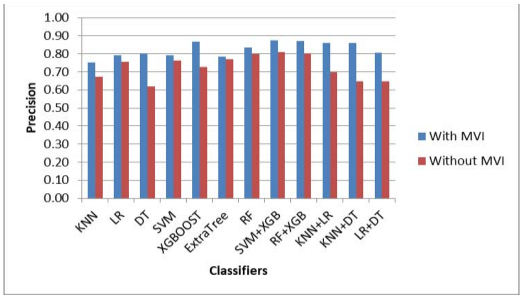
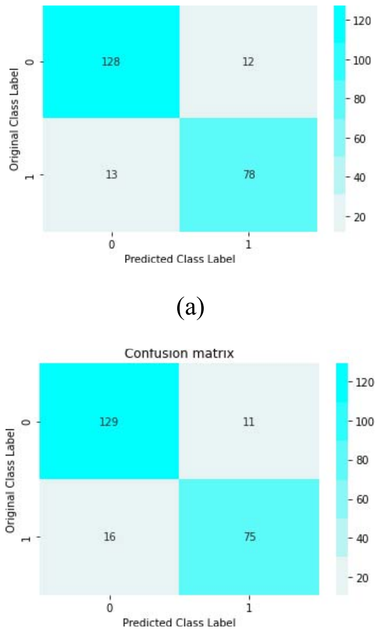


Figure 3. Precision Score of various classifier



(a)

(b)

Figure 4. Confusion Matrix (a) .XGBoost (b) SVC (XGB+RF)

Table III illustrates the accuracy score in the train-test split and 5-fold cross-validation. The Area Under the Curve (AUC) of various machine learning classifiers is shown in Fig 5.

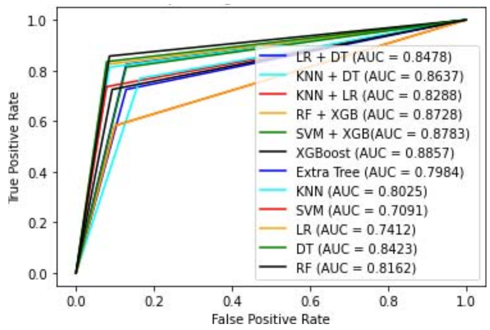


Figure 5. ROC curve for test data with.MVI

V. CONCLUSION

Early identification of diabetes is a big challenge in the current research. In our work, a framework capable of accurately predicting diabetes is presented. Soft voting ensemble method and gradient boosting model is used to predict the diabetes in Pima dataset. Twelve different ML algorithms, including KNN, SVM, DT, LR, RF and Ensemble Models to predict diabetes and evaluate performance on various measures like precision, recall, accuracy, F1 Score and AUC. This work also emphasizes missing value imputation. In both train-test split and k-fold cross-validation, overall accuracy improves significantly with missing value imputation. Among all the proposed models, the extreme gradient boosting classifier (XGBoost) and soft voting classifier by combining XBoost with RF is the most efficient and promising for predicting diabetes.

References

[1] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, 2021, doi: 10.1016/j.icte.2021.02.004.

[2] M. Ghaderi, M. A. Farahani, N. Hajiha, F. Ghaffari, and H. Haghani, "The role of smartphone-based education on the risk perception of type 2 diabetes in women with gestational diabetes," *Health Technol. (Berl).*, vol. 9, no. 5, pp. 829–837, 2019, doi: 10.1007/s12553-019-00342-3

[3] S. Mandal, "New molecular biomarkers in precise diagnosis and therapy of Type 2 diabetes," *Health Technol. (Berl).*, vol. 10, no. 3, pp. 601–608, 2020, doi: 10.1007/s12553-019-00385-6.

[4] S. Mandal, "New molecular biomarkers in precise diagnosis and therapy of Type 2 diabetes," *Health Technol. (Berl).*, vol. 10, no. 3, pp. 601–608, 2020, doi: 10.1007/s12553-019-00385-6.

[5] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[6] C. Phaloprakarn and S. Tangjitgamol, "Risk score for predicting primary cesarean delivery in women with gestational diabetes mellitus," *BMC Pregnancy Childbirth*, vol. 20, no. 1, pp. 1–8, 2020, doi: 10.1186/s12884-020-03306-y.

[7] "https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-t reatment/drc-20355284." .

[8] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[9] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[10] H. I. Okagbue, P. I. Adamu, P. E. Oguntunde, E. C. M. Obasi, and O. A. Odetunmibi, "Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer," *Health Technol. (Berl).*, no. 0123456789, 2021, doi: 10.1007/s12553-021-00572-4.

[11] L. Lama *et al.*, "Machine Learning for Prediction of Diabetes Risk in Middle-aged Swedish people," *Heliyon*, p. e07419, 2021, doi: 10.1016/j.heliyon.2021.e07419.

[12] L. C. Gregor Stiglic, Fei Wang, Aziz Sheikh, "Development and validation of the type 2 diabetes mellitus 10-year risk score prediction models from survey data," *Prim. Care Diabetes*, vol. 15, no. 4, p. Pages 699-705, 2021.

[13] O. Simon LebechCichosz, Morten HasselstrømJensen, "Short- term prediction of future continuous glucose monitoring readings in type 1 diabetes: Development and validation of a neural network regression model," *Int. J. Med. Inform.*, vol. 151, 2021.

[14] K. Vizhi and A. Dash, "Diabetes prediction using machine learning," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6, pp. 2842– 2852, 2020, doi: 10.32628/cseit2173107.

[15] M. F. N. Muhammad Daniyal Baig, "Diabetes prediction using machine learning algorithms," *Lect. Notes Networks Syst.*, no. November, 2020, doi: 10.13140/RG.2.2.18158.64328.

[16] D. S. V. Soni, Mitushi, "Diabetes Prediction using Machine Learning Techniques," *Int. J. Eng. Res. Technol.*, vol. 9, no. 09, pp. 921–924, 2020, doi: 10.1007/978-981-33-6081-5_34.

[17] T. V. Varga, K. Niss, A. C. Estampador, C. B. Collin, and P. L. Moseley, "Association is not prediction: A landscape of confused reporting in diabetes – A systematic review," *Diabetes Res. Clin. Pract.*, vol. 170, p. 108497, 2020, doi: 10.1016/j.diabres.2020.108497.

[18] T. P. Jaiswal, Varun, Anjli Negi, "A review on current advances in machine learning based diabetes prediction," *Prim. Care Diabetes*, vol. 15, no. 3, pp. 435–443, 2021.

[19] K. Kalagotla, Satish Kumar, Suryakanth V.Gangashetty, "A novel stacking technique for prediction of diabetes," *Comput. Biol. Med.*, vol. 135, 2021.

[20] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Comput. Methods Programs Biomed.*, vol. 202, 2021, doi: 10.1016/j.cmpb.2021.105968.