JOURNAL OF DATA SCIENCE | Vol.2024:47
eISSN:2805-5160

Diabetes Classification Using a Framework Stacking of BiLSTM, Logistic Regression, and XGBoost

M. Rezqy Noor Ridha[1], Silvia Ratna[2], M. Muflih[3], Haldi Budiman[4], Usman Syapotro[5], Muhammad Hamdani[6]

[1, 2, 3, 4, 5, 6]Faculty of Information Technology, Islamic University of Kalimantan Muhammad Arsyad Al-Banjari, Indonesia

Email: rezqyridha@gmail.com[1], silvia.ratna@uniska-bjm.ac.id[2], mmuflihfti@uniska-bjm.ac.id[3], haldibudiman@uniska-bjm.ac.id[4], 05usman.syapotro.study@gmail.com[5], mhdhamdani.formal@gmail.com[6]

Abstract

Diabetes is a chronic condition that requires accurate and timely diagnosis for effective management and treatment. This study introduces an innovative approach to diabetes classification using a stacking framework that combines Bidirectional Long Short-Term Memory (BiLSTM), Logistic Regression, and XGBoost. The study employed an experimental approach by implementing the stacking framework. The two base models used were BiLSTM and Logistic Regression, with BiLSTM achieving an accuracy of 0.9935 and Logistic Regression reaching 0.9869. The stacking framework with XGBoost as the meta-learner achieved a perfect accuracy of 1.0. These findings demonstrate the potential of the stacking framework to improve diabetes classification performance compared to using individual models alone.

Keywords

Diabetes Classification, Stacking, Bidirectional Long Short-Term Memory (BiLSTM), Logistic Regression, Extreme Gradient Boosting (XGBoost)

Introduction

Millions of individuals worldwide are impacted by diabetes, making it a serious global health concern. Diabetes is becoming more common as a result of sedentary lifestyles, bad diets, and longer life expectancies. Reducing the consequences of diabetes, such renal failure, neuropathy, and cardiovascular disease, requires early diagnosis and efficient care. The creation of precise and effective diagnostic instruments has become crucial due to the chronic nature of diabetes and its effects on public health systems (Alaa Khaleel & Al-Bakry, 2023).

A hybrid deep learning model that combines a bidirectional long short-term memory (Bi-LSTM) network with a convolutional neural network (CNN) was presented in earlier study by Madan et al., 2022. The Pima Indians Diabetes dataset yielded an accuracy of 98.85% for this model. Furthermore, studies (Carpinteiro et al., 2023) indicated that the best models for predicting diabetes were Support Vector Machines (SVM), Multilayer Perception (MLP), and

Submission: 20 October 2024; **Acceptance:** 21 November 2024

Gradient Boost Machine (GBM). With an accuracy of 80.0%, SVM, MLP, and GBM marginally surpassed each other with an accuracy of 80.2%, 80.4%, and 80.4%, respectively.

Novelty of this research lies in the innovative application of a stacking framework that strategically combines the strengths of diverse machine learning models for enhanced diabetes classification. This framework integrates the sequential learning capabilities of BiLSTM, the interpretability of Logistic Regression, and the robust ensemble learning of XGBoost. By leveraging the unique advantages of each model while mitigating their individual limitations, this approach aims to improve prediction accuracy, offering a potentially more effective solution for diabetes classification compared to traditional single-model methods.

In the following sections, we will delve into the methodological intricacies of our proposed stacking framework, detailing the data preprocessing steps, model training procedures, and evaluation metrics employed. The results and discussion section will present a comprehensive analysis of our experimental findings, comparing the performance of the stacked model against individual base models and other state-of-the-art approaches. Finally, the conclusions and recommendations section will summarize the key contributions of this research, highlighting its potential implications for diabetes prediction and management, and suggesting avenues for future exploration.

Methodology

This study introduces a stacking-based ensemble framework combining Bidirectional Long Short-Term Memory (BiLSTM), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost) to classify diabetes. The methodology employs an experimental approach, consisting of four main phases:

Data Preprocessing

First, the data is loaded from a CSV file. Then, the features and target are separated, followed by feature scaling using StandardScaler to ensure all features have the same scale. Next, the data is split into training (80%) and testing (20%) sets. Finally, the training and testing data are reshaped to fit the format required by the BiLSTM model.

BiLSTM Model

A BiLSTM model with two bidirectional LSTM layers and a Dense layer for multi-class classification has been constructed. To prevent overfitting, BatchNormalization and Dropout have been employed. The model is then compiled using the 'adam' optimizer, 'sparse_categorical_crossentropy' loss function, and 'accuracy' evaluation metric. Furthermore, EarlyStopping is utilized during training to halt the process if performance on the validation set does not improve. (Xiaoyan & Raga, 2023) conducted research on sentiment classification of mixed-language Chinese text comments, involving both long and short texts, using the BiLSTM-Attention model.

Logistic Regression Model

A Logistic Regression model is initialized with max_iter =1000 to set the maximum number of training iterations, multi_class='multinomial' to indicate a multi-class classification problem, and solver='lbfgs' to employ an efficient optimization algorithm. Subsequently, the model is trained using the fit() method with training data X_train (features) and y_train (target), where it learns the relationship between features and target to make future predictions. (Shah et al., 2020) conducted research on BBC news text classification using machine learning algorithms, comparing the performance of three classification algorithms: logistic regression, random forest, and K-nearest neighbor (KNN).

Stacking Extreme Gradient Boosting (XGBoost)

In the stacking stage, the probability predictions from the BiLSTM and Logistic Regression models are combined to create new features. Next, an XGBoost model is employed as a meta-model. This meta-model is responsible for learning and combining the predictions from both base models (BiLSTM and Logistic Regression), aiming to produce a more accurate final prediction. The XGBoost model is then trained using these newly created combined features. (Bakasa & Viriri, 2023) proposed a medical image classification model called Stacked Ensemble Deep Learning (SEDL) for pancreatic cancer classification using computed tomography (CT) scans, utilizing the XGBoost model as a strong learner or meta-model to produce the final classification.

Evaluation

Using a variety of criteria, such as accuracy, precision, recall, and F1-score, the efficacy of the individual models and the stacking ensemble was evaluated on the test dataset. These metrics offer a thorough assessment of classification performance, particularly with regard to maintaining the proper ratio of false positives to false negatives. Furthermore, in order to investigate categorization mistakes further, a confusion matrix was created. Every performance assessment was carried out utilizing the Scikit-learn toolkit.

Results and Discussion

The experimental results of diabetes classification using the XGBoost stacking framework are as follows:

Table 1 Accuracy Results

Model	Dataset Split	Accuracy
BiLSTM	80% Train, 20% Test	0.99
Logistic Regression	80% Train, 20% Test	0.98
Framework Stacking (XGBoost)	**80% Train, 20% Test**	**1.0**

Accuracy results presented in Table 1 demonstrate that all three models employed BiLSTM, Logistic Regression, and Stacking (with XGBoost as the meta-learner) exhibit exceptional performance. BiLSTM achieves a high accuracy of 0.9934, showcasing its capability to capture complex patterns and long-term dependencies within the data, making it particularly valuable for tasks such as natural language processing and time series analysis. Despite its relative simplicity, Logistic Regression also delivers impressive performance with an accuracy of 0.9869, suggesting a predominantly linear relationship between the features and the target variable. Meanwhile, Stacking with XGBoost as the meta-learner achieves a perfect accuracy of 1.0, underscoring the power of ensemble learning in combining predictions from multiple models to yield superior results. XGBoost effectively learns the optimal way to integrate the predictions from both BiLSTM and Logistic Regression, thereby attaining peak performance.

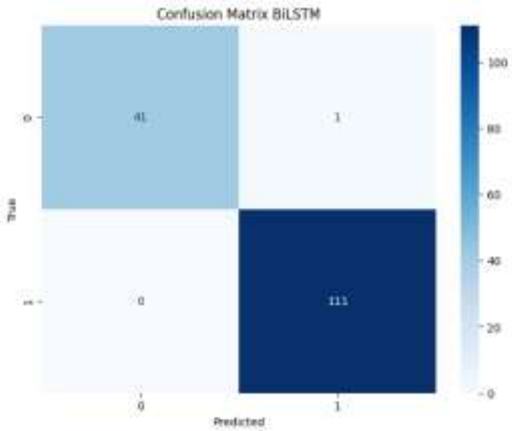

Figure 1 BiLSTM Confusion Matrix

Model BiLSTM correctly predicted 111 true positives, indicating data points with a positive label (1) accurately classified as positive. Furthermore, 41 true negatives represent data points with a negative label (0) correctly predicted as negative. One false positive exists, signifying a data point with a negative label incorrectly classified as positive (Type I error). Importantly, zero false negatives were observed, indicating no data points with a positive label were misclassified as negative.

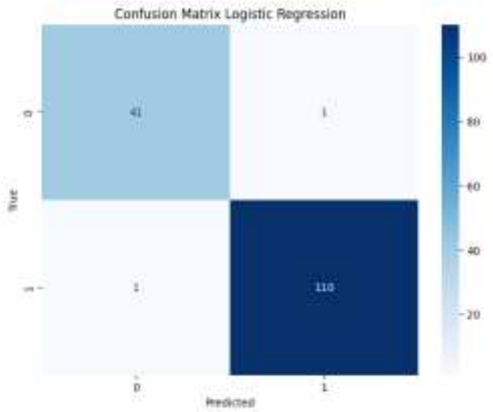

Figure 2 Logistic Regression Confusion Matrix

Model Logistic Regression accurately predicted 110 true positives, signifying observations with a positive label (1) correctly classified as positive. Additionally, 41 true negatives denote observations with a negative label (0) correctly predicted as negative. One false positive exists, indicating an observation with a negative label incorrectly classified as positive (a Type I error). Lastly, one false negative was observed, signifying an observation with a positive label incorrectly predicted as negative.

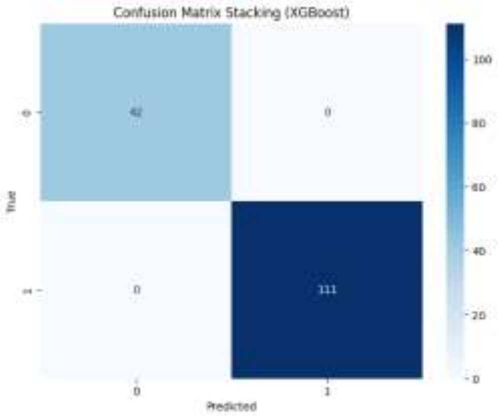

Figure 3 Stacking XGBoost Confusion Matrix

Stacking XGBoost achieved 111 true positives, signifying observations with a positive label (1) correctly classified as positive. In addition, 42 true negatives represent observations with a negative label (0) correctly predicted as negative. Notably, zero false positives were observed, indicating no observations with a negative label were incorrectly predicted as positive (Type I error). Finally, zero false negatives were recorded, signifying that no observations with a positive label were incorrectly predicted as negative.

Conclusion

This study successfully demonstrated the effectiveness of a stacking framework combining BiLSTM, Logistic Regression, and XGBoost for diabetes classification. The stacking ensemble, leveraging the strengths of each individual model, achieved a perfect accuracy of 1.0, outperforming both BiLSTM (0.9934) and Logistic Regression (0.9869). This highlights the potential of ensemble learning in enhancing prediction accuracy for complex medical diagnoses like diabetes.

Recommendation

In order to further enhance performance and generalizability, future studies might investigate using this approach to bigger and more varied datasets and maybe adding new models or feature engineering techniques. The knowledge gathered from this research helps with the continuous creation of sophisticated machine learning instruments for diabetes management and prediction, which in turn helps with early diagnosis and individualized treatment plans.

References

Alaa Khaleel, F., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*, *80*(xxxx), 3200–3203. https://doi.org/10.1016/j.matpr.2021.07.196

Bakasa, W., & Viriri, S. (2023). Stacked ensemble deep learning for pancreas cancer classification using extreme gradient boosting. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1232640

Carpinteiro, C., Lopes, J., Abelha, A., & Santos, M. F. (2023). A Comparative Study of Classification Algorithms for Early Detection of Diabetes. *Procedia Computer Science*, *220*, 868–873. https://doi.org/10.1016/j.procs.2023.03.117

Madan, P., Singh, V., Chaudhari, V., Albagory, Y., Dumka, A., Singh, R., Gehlot, A., Rashid, M., Alshamrani, S. S., & Alghamdi, A. S. (2022). An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment. *Applied Sciences (Switzerland)*, *12*(8). https://doi.org/10.3390/app12083989

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1). https://doi.org/10.1007/s41133-020-00032-0

Xiaoyan, L., & Raga, R. C. (2023). BiLSTM Model With Attention Mechanism for Sentiment Classification on Chinese Mixed Text Comments. *IEEE Access*, *11*(March), 26199–26210. https://doi.org/10.1109/ACCESS.2023.3255990