Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 8 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i08.48387

PAPER

High Performance for Predicting Diabetic Nephropathy Using Stacking Regression of Ensemble Learning Method

Lailil Muflikhah[1](✉), Amira Ghina Nurfansepta[1], Fitra Abdurrachman Bachtiar[1], Dian Eka Ratnawati[2]

[1]Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

[2]Department of Information System, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

lailil@ub.ac.id

ABSTRACT

Diabetes may lead to several problems, one of the most prevalent and deadly of which is diabetic nephropathy. Therefore, the condition represents a significant threat to one's health since it has the potential to cause irreversible harm to the kidneys' ability to operate. A significant portion of the research that is being conducted now is focused on determining how accurately diabetic people may be predicted to develop kidney illness. Considering this, the research suggests a regression stacking approach for predicting albumin levels. These albumin values will serve as a reference for the incidence of diabetic nephropathy disease. They will be derived from the medical records of patients. The utilization of stacking regression from three different ensemble approaches, using Random Forest and CatBoost regressors, while the Huber algorithm is used as a meta-learner. The accuracy with which the combination of parameters that are employed is determined is a significant factor. It contributes to the high degree of performance that the ensemble approach achieves. Therefore, in this investigation, a grid search was carried out to tune the hyperparameters of both regressor models. We evaluated the performance of the proposed model using accuracy, MAPE, RMSE, and MSE values. The experimental findings demonstrate great performance. Three selected variables including quantitative UACR, semi-quantitative UACR, and urinary creatinine, achieved high performance. Overall, the performance obtained an accuracy rate of more than 98% with an error rate (MAPE, RMSE, and MSE values) of less than 1%. In conclusion, the stack regressor model can be implemented to predict diabetic nephropathy using clinical datasets.

KEYWORDS

diabetic nephropathy, ensemble method, machine learning, stack regression

Muflikhah, L., Nurfansepta, A.G., Bachtiar, F.A., Ratnawati, D.E. (2024). High Performance for Predicting Diabetic Nephropathy Using Stacking Regression of Ensemble Learning Method. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(8), pp. 149–164. https://doi.org/10.3991/ijoe.v20i08.48387

Article submitted 2024-02-05. Revision uploaded 2024-03-25. Final acceptance 2024-03-25.

1 INTRODUCTION

Diabetes is the term used to describe hyperglycemia resulting from inadequate production and/or resistance to insulin. According to the International Diabetes Federation (IDF), the global population of persons aged 20 to 79 with diabetes is projected to reach 537 million by 2030 and 784 million by 2045 [1]. Currently, diabetes ranks as the seventh most prevalent cause of mortality on a global scale. Most diabetes individuals in both developed and underdeveloped nations experience microvascular problems. The problems, such as diabetic nephropathy, retinopathy, and neuropathies, contribute significantly to the morbidity and mortality of diabetic patients [2].

Diabetic nephropathy (DN) is a serious long-term consequence of diabetes that affects the small blood vessels in the kidneys. It is a leading cause of advanced renal disease and is linked to a higher risk of heart problems [3]. Nephropathy poses a significant health obstacle. It is present in a maximum of 50% of individuals with diabetes [4]. Nephropathy is a medical condition marked by long-lasting albuminuria and a gradual deterioration of kidney function. Nephropathy is mostly caused by chronic hyperglycemia and hypertension. Typically, it is recommended to do an annual screening for microalbuminuria. This should begin 5 years after the diagnosis of type 1 diabetes, at the time of diagnosis, and annually thereafter for type 2 diabetes [5]. Nephropathy is a diverse condition marked by the occurrence of albuminuria in individuals with diabetes. Pathological albuminuria, as defined by the American Diabetes Association (ADA), is characterized by an albumin-to-creatinine ratio (ACR) over 30 mg/g. In individuals with type 1 diabetes mellitus, the occurrence of nephropathy is infrequent during the initial decade of the disease. However, the likelihood of developing nephropathy rises progressively between 10 and 20 years following diagnosis, with an annual increase of up to 3%. The regression of albuminuria remains a significant objective. Indeed, the incidence of nephropathy is increasing and necessitates proactive measures to prevent or rectify the illness.

The utilization of data-driven technology serves as the foundation for the digital healthcare revolution, offering a crucial chance to enhance personalized treatment and promote advancements in medical research [6]. Artificial intelligence (AI) is a widely utilized technology that has shown tremendous growth in the past decade. Machine learning (ML) is now the most well-recognized division of AI. Most systems that employ machine learning techniques utilize them to do predictive analysis. AI and ML are scientific disciplines that employ computer algorithms to acquire knowledge from data, facilitate the detection of patterns within data, and generate predictions. Utilizing machine learning for predictive analysis can offer a viable answer to the challenge [7], [8], [9].

Previous research has conducted many experiments on utilizing machine learning to predict nephropathy in diabetic individuals. In 2021, comparison research was conducted on machine learning algorithms to predict outcomes for a group of 133 patients. The current characteristics are integrated into six machine learning-based models, specifically LDA, SVM, Logistic Regression, KNN, Naïve Bayes, and simulated neural networks. When PCA is used with SVM-RBF, the resulting accuracy is 88.7% and the AUC is 0.91 at a cutoff of 0.96 PCA [10]. In 2022, research utilized extensive data from 23 diabetes facilities in Italy. The study centered on utilizing the XGBoost model to forecast several consequences associated with diabetes. After considering all the complexities, the model demonstrated an accuracy rate of over 70%, with a specific accuracy of 89.7% for nephropathy. Additionally, the model achieved an Area Under the Curve (AUC) of more than 80% [11].

The difference from previous research in this area relates to the accuracy of predicting the occurrence of nephropathy related to albumin levels based on clinical

data of diabetic patients at Saiful Anwar hospital, Malang. Considering this, the aim of this study is to develop a reliable method to predict albumin levels, which can then be used by medical professionals to determine whether kidney failure (nephropathy) occurs in such patients. We propose to study the prediction of nephropathy based on albumin values (continuous numerical data). Nephropathy is said to occur if the albumin value in the blood is >300, otherwise (albumin value ≤300), it is considered normal.

In addition, the accurate setting of parameter values and the imbalance of data distribution in each class lowered the performance of various machine learning techniques. Therefore, this study proposes collecting the regression of the ensemble learning method by adjusting the hyperparameter. By combining the strength of various classifier models, high predictive performance is expected.

2 LITERATURE REVIEW

Numerous studies have investigated the prediction of Diabetic Nephropathy (DN) risk through the integration of machine learning and feature extraction methodologies. Maniruzzaman et al. (2021) studied to compare machine learning (ML) approaches for the identification of risk factors for nephritis. In this work, the researcher utilized principal component analysis (PCA) to identify features in the DN risk prediction model. The method used combines machine learning (ML) with PCA to determine the optimal cutoff value that yields the highest level of precision. Six machine learning methods are then used to apply the best features: CNN, LDA, SVM, Logistic Regression, Naïve Bayes, and replicated neural networks. By using a leave-one-out cross-validation process, the performance of machine learning-based engineering was evaluated and compared using accuracy and area under the curve (AUC) metrics. The study's findings show that when a Support Vector Machine with Radial Basis Function (SVM-RBF) is coupled with Principal Component Analysis (PCA) at a threshold of 0.96 PCA, the best degree of accuracy (88.7%) and Area Under the Curve (AUC) (0.91) are attained. Also, research on disease detection using the ensemble method of machine learning achieved a high-performance accuracy rate [12], [13], [14].

Another research by Nicolucci et al. (2022) is to use big data and machine learning to create a predictive model for diabetes complications retrieved from electronic medical records. Twenty-three centers used data from 147,664 patients during fifteen years. Five more facilities carry out external validation. Nephropathy, diabetic neuropathy, ophthalmic, cardiovascular, cerebrovascular, and peripheral vascular disorders form the six categories of problems addressed in this study. Task 1 uses the XGBoost algorithm to forecast the probability that difficulties will arise during the next five years. Following this, the technique for predicting concerns is partitioned into two temporal phases: the initial phase, anticipated to occur in two years, and the succeeding phase, expected to transpire within three to five years. The model was evaluated using the ROC curve (AUC) and accuracy rate. It achieved an accuracy of 70% and AUC of 0.8-0.97. On the other hand, task 2, the accuracy of each prediction model is above 70%, and its associated AUC (Area Under the Curve) value topped 0.85. For peripheral vascular disease and nephropathy, respectively, the percentage of sensitivity in predicting the early start of problems is 83.2% and 88.5%.

Furthermore, a study on stack regression was implemented by Gadgil et al. (2023) to solve the problem of Customer Lifetime Value prediction. That study was compared to the conventional method and showed improvement in the stacked regression methods [9][15]. The dataset included in this investigation is

called “Online Retail II,” and it consists of the transactions carried out by an online retailer situated in the United Kingdom from December 1, 2009 to December 30, 2011. The employed layered regression approach incorporates meta-learning. Using this method, predictions from many basic models—like XGBoost, Random Forest, and elasticNet—are integrated. After that, elasticNet, the meta-model, receives the combined outputs of these basic models as input. The outcomes show that the suggested layered regression strategy performs better than the selected base model. As assessment measures, the lowest Mean Absolute Error (MAE) value of 0.82 and the second-lowest Root Mean Square Error (RMSE) value of 1.37 are used. The research undertaken by Nsenge Mpia et al. (2023) aims to predict the duration of hospitalization for individuals receiving medical care in semi-urban facilities. Additionally, the study assesses the efficacy of multiple models in this regard. The information utilized was collected at Clinique La Lumière, which is in the Butembo municipality of the North Kivu Province in the Democratic Republic of the Congo. The dataset comprises 838 samples collected between January 2, 2010, and August 25, 2021. The methodology implemented in this research is layered regression. The research utilizes a compilation of six unique regression models to develop a model: Random Forest, Extra Trees, Decision Tree, XGBoost, Light GBM, and SVR. Individual models performed less well than stacked regression, which achieved an accuracy of 91%, MAE of 0.12, MSE of 0.14, and RMSE of 0.37.

Using predictive analysis, the study by Cui, et al. (2023) attempts to gain a comprehensive comprehension of the rapid charging behavior of electric vehicle consumers at stations. The dataset utilized in this study comprises more than 220,000 authentic charging records provided by consumers of electric vehicles at public charging stations. Ensemble learning is implemented in this paper using stacked regression models, including Ridge, Random Forest, and LightGBM. Each fundamental model is initially trained using K-Fold and GridSearch techniques to produce optimal prediction models [16].

3 RESEARCH METHOD

In this research, patients who were diagnosed with diabetics were used to compile a dataset. People who have been diagnosed with diabetic nephropathy are referred to the certain characteristic. There are 164 individuals suffering from nephropathy and 336 patients living with normal kidney function that is included in this dataset, which has a total of 500 rows of patient data. This study employed 80% (395) of the data for training and 20% (99) for testing.

3.1 The proposed method

In general, this research started with clinical data acquisition. This research used secondary data from the Information System of medical records in Saiful Anwar Hospital, Malang. Then, it is applied preprocessing to improve the data quality, including cleansing data, handling missing values, centering data values, and feature selection by getting the correlation of multivariate data. Then, the data is split into training and testing. The stacking regression is implemented from the ensemble method, including Random Forest, CatBoost, and Huber Algorithm using data training. By tuning the hyperparameter for the regressor, the result for evaluation was achieved. The steps are detailed in Figure 1.

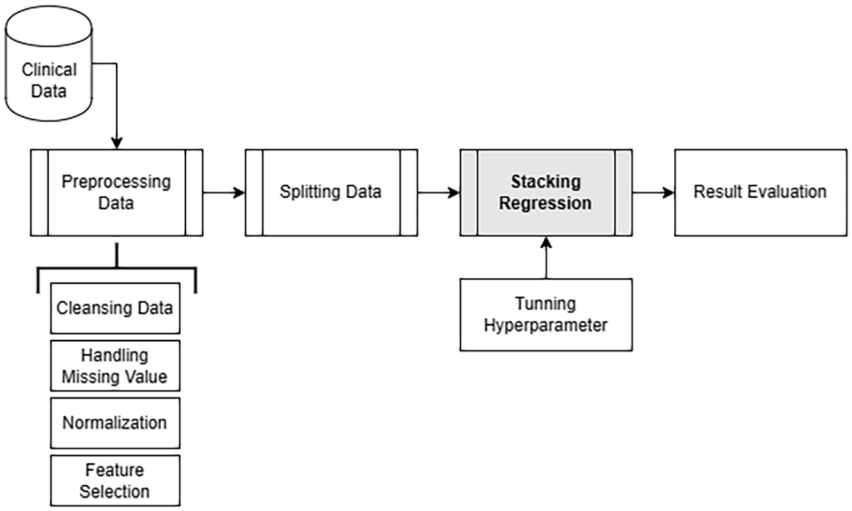

Fig. 1. General proposed method

Initially, we gather pertinent clinical data from patients diagnosed with diabetes. The data may encompass diverse attributes, such as age, creatinine level, creatinine urine, and other relevant health indicators. The objective of the class is to determine the albumin score. This biomarker is essential for evaluating kidney function. Next, we improve the data quality by preprocessing it, which involves addressing missing values and reducing attribute value variation by normalization on a consistent scale. To obtain information about the link between data, we do correlation analysis. We choose the attribute with a strong correlation value for the subsequent modeling phase. Subsequently, we divided the dataset into separate training and testing sets. The training set is utilized for model training, whereas the testing set is employed to evaluate the models' generalization performance.

In the classifier models, such as the one shown in Figure 2, we utilized the training data to create a stacking ensemble learning approach. For prediction in the first level, we have selected the Random Forest and CatBoost algorithms as regressors. The hyperparameters of these regressors are tuned using the grid search approach. Subsequently, we employ the Huber algorithm as a meta-learner to make the ultimate prediction. Finally, we evaluate their performance by assessing accuracy, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE).

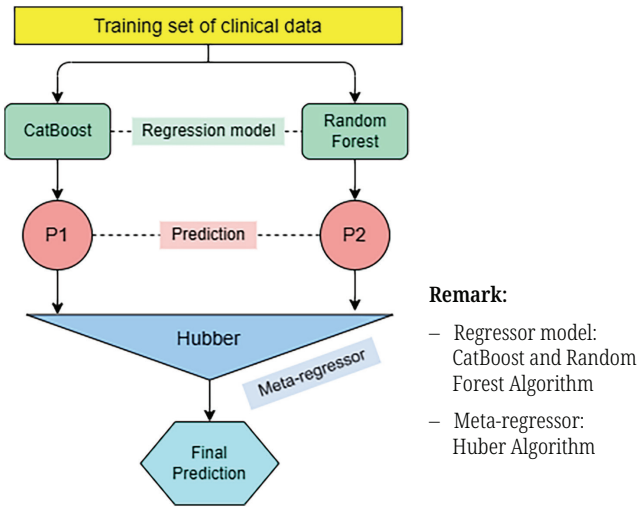

Fig. 2. The architecture of the stacked regression model

3.2 Prediction

Prediction models are designed to assist healthcare professionals and patients in making decisions regarding the use of diagnostic tests, starting, or stopping treatment, or making lifestyle changes. While not a substitute for clinical treatment, these models can provide objective data about a person's disease risk and avoid some of the biases common in clinical decision-making [17]. Moreover, the information generated in healthcare is growing very fast and outpacing the capacity of human cognition to adequately manage it.

Risk prediction models use patient characteristics to estimate the probability that a certain outcome exists or will occur within a certain timeframe. Predictions are obtained from the results of regression methods. The regression method is a measurement method based on the relationship between one or many other variables [18]. This regression method attempts to correlate the existing dependent variables with the given independent variables. Regression analysis will produce a regression model. In this research, we applied regression method to predict albumin value level as threshold for determining nephropathy disease.

Ensemble Method of Machine Learning. The ensemble technique is used to improve the performance of prediction purposes. Ensemble learning tends to have higher performance evaluation than single model trained with a particular technique to reduce the overall error rate and improve model performance [7], [19]. Several models can be trained on different parameter sets to create multiple decision boundaries on randomly selected data points as training data. Therefore, by using ensemble learning techniques, these problems can be addressed and reduced by training multiple algorithms, and the results can be combined to obtain near-optimal results.

The ensemble method can be done by combining the outputs of multiple base models in several ways or using multiple methods to select the best base model. The stacking technique is one of the ensemble techniques where a meta-learning model is used to integrate the outputs of the base models. If the final decision part is a linear model, stacking is often referred to as meta-learner [20]. The concept of stacking regression was originally given by Wolpert in 1992. In this technique, the dataset is randomly divided into J equal parts. For the *j*th cross-validation, one set is used for testing and the rest for training. By training and testing this subset of pairs, the predictions of different learning models are obtained and used as meta-data to build meta-models. The meta-model makes the final prediction of the other models.

Huber Algorithm. As one of the milestones of robust statistics, Huber regression plays an important role in robust inference and estimation. It has also been used in a wide variety of applications in machine learning [21]. The regression paradigm associated with the Huber loss is referred to as the Huber regression estimator. The Huber loss function can be defined as stated in Equations (1) and (2).

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{m} \varnothing(y_i - x_i^T \beta) \tag{1}$$

$$\varnothing(u) = \begin{cases} u^2, if \left|u\right| \le M \\ 2Mu - M^2, if \left|u\right| > M \end{cases} \tag{2}$$

The setting parameters of the Huber algorithm for diabetic nephropathy prediction include an alpha value is 0.1. The Huber regression function controls loss by applying a margin of 1.0 and limiting iterations to 100 for maximum control in recovery training.

Random Forest Algorithm. Random Forest consists of a combination of decision trees. It improves the classification performance of single-tree classifiers by combining bootstrap aggregation [7]. It is also called the bagging method and randomization in the selection of data partitioning nodes in decision tree construction [22]. A decision tree with M leaves divides the feature region into M regions R_m, $1 \le m \le M$. For each tree, a Random Forest can be defined as in Equation (3) and (4).

$$f(x) = \sum_{m=1}^{M} c_m \prod (x, R_m) \tag{3}$$

$$\prod (x, R_m) = \begin{cases} 1, \ if \ x \in R_m \\ 0, otherwise \end{cases} \tag{4}$$

The study proposed an algorithm controlling the maximum depth in each tree and the number of ensemble Random Forest trees using a parameter of 100.

CatBoost Algorithm. CatBoost was developed by researchers and engineers at Yandex, a Russian company, in 2017, and released its code in April. CatBoost is the latest algorithm in the open-source field in the world today and surpasses other algorithms, such as XGBoost and LightGBM in terms of performance. CatBoost is a boosting algorithm and effectively solves the problems of gradient bias and prediction drift. It avoids overfitting, and improves calculation accuracy and generalization ability [23]. This method reduces the effects of noise and categorical low frequency on data distribution as in Equation (5).

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1} \left[X_{\sigma j,k} = X_{\sigma p,k} \right] . Y_{\sigma_j} + a.p}{\sum_{j=1}^{p-1} \left[X_{\sigma j,k} = X_{\sigma p,k} \right] + a} \tag{5}$$

Where p is the previous term added and a is a weighting factor greater than 0.

CatBoost is a complex algorithm used in machine learning tasks, based on gradient-boosted decision trees. It handles categorical features and missing values, using formulas to calculate loss functions and optimized metrics. CatBoost's score function measures the quality of gradient approximation, providing functions like L2, Cosine, NewtonL2, and NewtonCosine. These functions can improve model results and can't be used with loss guide tree growth policy. This research establishes parameters for controlling sedentary behavior in CatBoost assemblies, including depth of 4 and iterations of 1000, and a learning rate of 0.01, ensuring significant body weight change monitoring during each iteration.

3.3 Tunning grid search hyperparameter

GridSearchCV is a machine-learning technique that uses hyperparameter tuning to find the best combination for a given algorithm. It is applied exclusively to the training dataset, ensuring its integrity for unbiased evaluation.

The basic principle of the GridSearch method is to divide the grid into a certain range and traverse all points in the grid with parameter values C and d. Combined with the Cross Validation (CV) method, the parameter value method is obtained. Finally, the parameters C and s with the highest accuracy are determined as the best parameters [24]. Model optimization is performed by tuning the best hyperparameters generated from GridSearch. The main parameters optimized are the number of trees, the maximum number of features to be shared in child nodes, and the

level of the trees in each decision tree [25]. In this research, tuning hyperparameters is aimed at getting the best combination of the ensembled method's parameters. The following are the steps in using GridSearch:

a) To specify the model and parameters to be tuned.
b) To determine the range of values for each parameter.
c) To create a grid of all possible parameter combinations.
d) To train the model for each parameter combination using cross-validation.
e) To select the parameter combination that gives the best performance.

3.4 Performance evaluation

To evaluate the performance, the prediction models were then validated using several indicators, including Mean Absolute Error (MAE), Mean Square Error (MSE), and RMSE.

Mean Absolute Error (MAE). Mean Absolute Error (MAE) is a statistical tool that may be utilized in situations when an outlier is present in the data. According to Chicco et al. (2021), MAE does not adequately account for the fact that, due to the great number of penciled trainees (because of the presence of outliers), it results in the provision of a work schedule that is both standard and variable for the model [26]. This is how MAE may be defined in Equation (6).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|X_i - Y_i\right| \tag{6}$$

Mean Absolute Error (MSE). Mean Square Error (MSE) can be used if there are outliers that need to be detected. MSE is great for attributing greater weight to such points. If the model ends up producing one very poor prediction, then the squared part of the function magnifies the MSE as defined in Equation (7).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \tag{7}$$

Root Mean Square Error (RMSE). Root Mean Square Error (RMSE) is the square root result of the MSE value. The order of regression models based on MSE will be identical to the order of models based on RMSE. RMSE can be defined as in Equation (8).

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \tag{8}$$

4 RESULT AND DISCUSSION

This study employed 80% (395) of the data for training and 20% (99) for testing. We utilized a grid search to collect the training data for this study. The optimal parameters are subsequently determined. Following that, these parameters are utilized to train the model, which is subsequently put to the test. We assessed the effectiveness of combining CatBoost and Random Forest algorithms as regressor models. Then, Huber algorithm is implemented as the meta-regressor. The use of stacking regression in ensemble learning techniques is addressed to achieved high-performance diagnosis of diabetic nephropathy. Our findings indicate promising outcomes in

terms of prediction accuracy and model resilience. We discovered that the stack regression model outperformed the individual base models (CatBoost and Random Forest) and traditional regression techniques. The stack regression ensemble was more successful in recognizing the intricate patterns. CatBoost is known for its robust handling of categorical variables and gradient boosting framework. On the other hand, the Random Forest provided stability and generalization ability with its ensemble of decision trees and inherent resistance to overfitting.

This meta-regressor helped to improve the predictions made by the underlying models, which led to a more accurate and dependable final projection. In terms of model interpretability, the stack regression ensemble provided valuable information regarding the underlying relationships between predictor variables and the incidence of diabetic nephropathy. By merging the predictions of many models, the ensemble approach assisted in the diagnosis and treatment decision-making process by offering a comprehensive understanding of the factors underlying the disease. Our results show that stacking regression is an effective ensemble learning method for diagnosing diabetic nephropathy. Combined with the stability of the Huber algorithm as the meta-regressor and the complementary capabilities of Random Forest and CatBoost as regressor models, we were able to obtain excellent prediction performance and improved interpretability of the model. This has enhanced clinical decision support and patient outcomes when it comes to the diagnosis of diabetic nephropathy.

GridSearchCV evaluates each hyperparameter combination using cross-validation techniques, mitigating the risk of overfitting. The best hyperparameter combination is identified, representing the model's optimal configuration. The trained model is then used to train the final model, fine-tuning it to achieve the best possible performance based on the training data. Finally, the trained model is evaluated on the testing dataset, providing an unbiased assessment of its effectiveness in real-world scenarios. In the Huber regression model, GridSearch was employed to determine the optimal parameters among a selection, including 'alpha': [0.1, 0.5, 1.0], 'epsilon': [1.0, 1.5, 2.0], and 'max_iter': [100, 200, 300]. The best parameter values obtained were 'alpha': 0.1, 'epsilon': 1.0, and 'max_iter': 100. These values include alpha: 0.1, epsilon: 1.0, and max_iter: 100. Then, they were applied in tuning the model and the result is shown in Figure, 3.

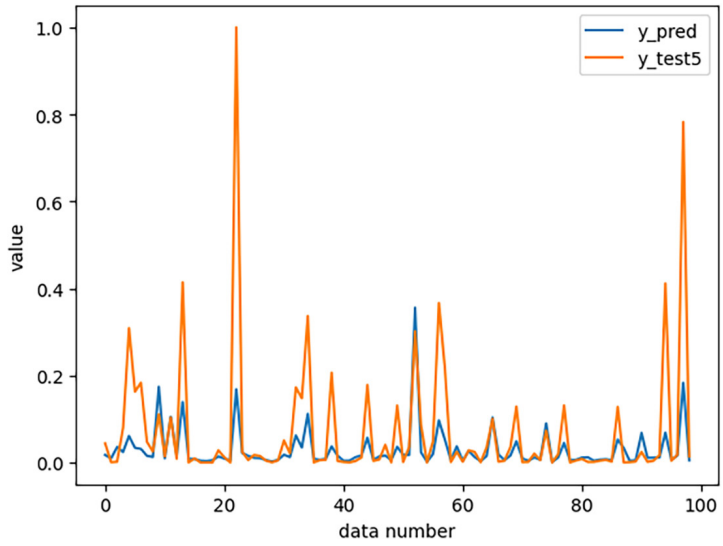

Fig. 3. Diabetic nephropathy prediction using huber algorithm

Conversely, the optimal parameters for DN prediction using Random Forest with tuning hyperparameters via Grid Search. They were identified from a set of

alternatives, which included the values 'n_estimators' (100, 200, 300) and'max_depth' (None, 10, 20). The optimal value of the parameters obtained was 'n_estimators': 100 and'max_depth', and None. By tuning of these parameters (max_depth: None, n_estimators: 100), the model was performed as in Figure 4.

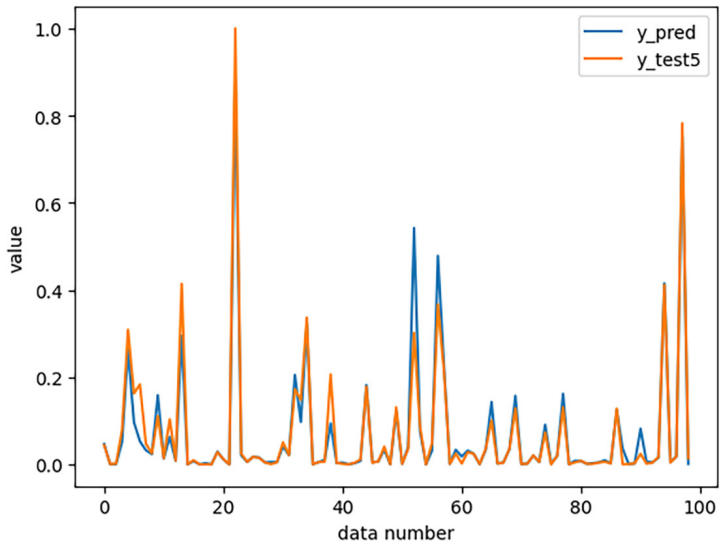

Fig. 4. Diabetic nephropathy prediction using random forest algorithm

The parameters are optimized through GridSearch in CatBoost for Diabetic Nephropathy prediction. The optimal values for 'iterations' (ranging from 100 to 500 to 1000), 'depth' (from 4 to 6), and 'learning_rate' (from 0.1 to 0.2) were determined through hyperparameter tuning via GridSearch. The optimal configuration of the parameters was as follows: 'depth' = 4, 'iterations' = 1000, and 'learning_rate' = 0.01. Then, the tuning of the parameters (depth: 4, iterations: 1000, learning_rate: 0.01), the model was produced as shown in Figure 5.

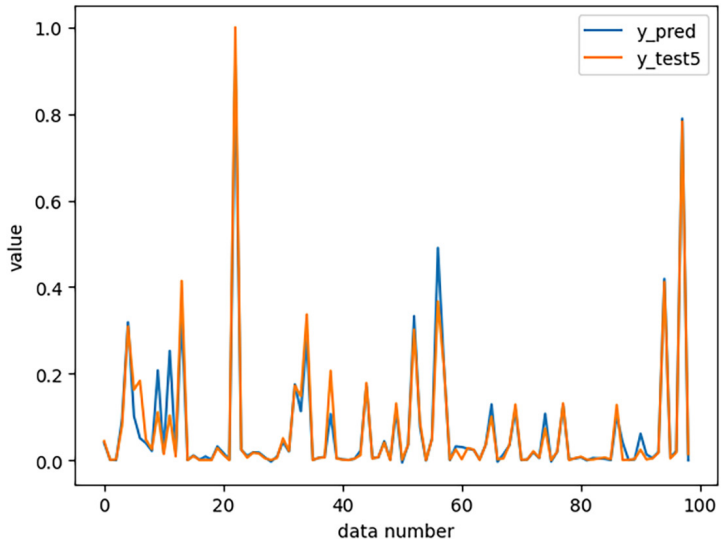

Fig. 5. Diabetic nephropathy prediction using catboost algorithm

Finally, a stacked regressor was utilized for the purpose of Diabetic Nephropathy prediction. This regressor included both CatBoost and Random Forest as regressors, and Huber was utilized as the meta regressor component. Figure 6 illustrates the outcomes of the model.

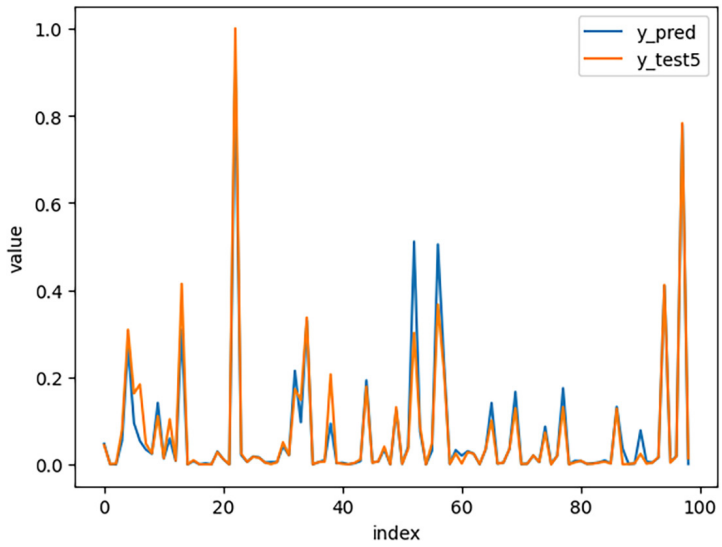

Fig. 6. Diabetic nephropathy prediction using stacking regression algorithm

The correlations among variables, such as Patient ID, Semi-quantitative UACR, Quantitative UACR, and Urinary Creatinine, were thoroughly investigated in the analysis of clinical data from diabetes patients. To find out if these factors were dependent on each other, this study analysis multivariate correlations. Significant prediction possibilities within the dataset were suggested by the strong correlation coefficients that were obtained. Out of all the variables, Semi-quantitative UACR, Quantitative UACR, and Urinary Creatinine had the strongest association values as shown in Figure 7. Their possible significance as diabetes patient data predictors was highlighted by these results.

Fig. 7. Multivariate correlation in diabetic nephropathy prediction

Subsequently, models were constructed to identify predictive determinants based on test data. As a result, predictions achieved various levels of high accuracy and had a low error rate. Comprehensive knowledge of prediction capabilities and performance indicators of the model being evaluated is shown in Tables 1 to 6.

Table 1. Performance of error rate for model (1): patient ID, semi-quantitative UACR, quantitative UACR, and Urinary Creatinine

Model	Duration	MAE	MSE	RMSE	RMSE (CV)	MAPE	Accuracy (1-MAPE)
RandomForestRegressor	0.564	0.021	0.003	0.055	0.071	0.020	0.980
StackingRegressor	4.896	0.020	0.003	0.051	0.072	0.019	0.981
CatBoostRegressor	5.320	0.024	0.004	0.065	0.064	0.187	0.813
Huber	0.036	0.066	0.026	0.160	0.151	0.091	0.909

Table 2. Performance of error rate for model (2): age, semi-quantitative UACR, quantitative UACR, and Urinary Creatinine

Model	Duration	MAE	MSE	RMSE	RMSE (CV)	MAPE	Accuracy (1-MAPE)
RandomForestRegressor	0.461	0.021	0.003	0.055	0.071	0.020	0.980
StackingRegressor	4.259	0.020	0.003	0.051	0.072	0.019	0.981
CatBoostRegressor	2.749	0.024	0.004	0.065	0.064	0.187	0.813
Huber	0.026	0.066	0.026	0.160	0.151	0.091	0.909

Table 3. Performance of error rate for model (3): patient ID, age, semi-quantitative UACR, quantitative UACR, and Urinary Creatinine

Model	Duration	MAE	MSE	RMSE	RMSE (CV)	MAPE	Accuracy (1-MAPE)
StackingRegressor	4.069	0.021	0.003	0.050	0.072	0.021	0.979
RandomForestRegressor	0.635	0.022	0.003	0.053	0.073	0.022	0.978
CatBoostRegressor	1.629	0.025	0.004	0.063	0.070	0.046	0.954
Huber	0.031	0.067	0.026	0.162	0.147	0.062	0.938

Table 4. Performance for model (4): semi-quantitative UACR, quantitative UACR, and Urinary Creatinine

Model	Duration	MAE	MSE	RMSE	RMSE (CV)	MAPE	Accuracy (1-MAPE)
RandomForestRegressor	0.372	0.019	0.002	0.046	0.069	0.017	0.983
StackingRegressor	4.186	0.018	0.002	0.043	0.069	0.016	0.984
CatBoostRegressor	0.817	0.022	0.004	0.059	0.069	0.021	0.979
Huber	0.039	0.054	0.020	0.140	0.228	0.054	0.946

Table 5. Performance for model (5): quantitative UACR and Urinary Creatinine

Model	Duration	MAE	MSE	RMSE	RMSE (CV)	MAPE	Accuracy (1-MAPE)
CatBoostRegressor	0.778	0.014	0.001	0.032	0.062	0.028	0.972
StackingRegressor	4.146	0.016	0.001	0.037	0.068	0.016	0.984
RandomForestRegressor	0.365	0.016	0.002	0.039	0.068	0.016	0.984
Huber	0.023	0.051	0.016	0.128	0.228	0.047	0.953

Table 6. Performance of error rate for five models in stacking regressor model

Model	Duration	MAE	MSE	RMSE	RMSE (CV)	MAPE	Accuracy (1-MAPE)
StackingRegressor (5)	4.146	0.016	0.001	0.037	0.068	0.016	0.984
StackingRegressor (4)	4.186	0.018	0.002	0.043	0.069	0.016	0.984
StackingRegressor (2)	4.259	0.021	0.002	0.048	0.069	0.018	0.982
StackingRegressor (1)	4.896	0.020	0.003	0.051	0.072	0.019	0.981
StackingRegressor (3)	5.350	0.021	0.003	0.050	0.072	0.021	0.979

The main contribution of our study lies in the introduction of the accumulation regressor as a new approach to addressing the predictive task for diabetic nephropathy. This study aims to explore the potential benefits of combining several regression models through accumulations to improve prediction performance in this domain. Based on the experimental result, the stacking regressor method does not show a significant improvement compared to Random Forest in terms of predictive outcomes while the longer training time and inference are associated to accumulation regressors compared to Random Forest.

5 CONCLUSION AND FUTURE WORK

This research provides strong evidence that ensemble learning methods – in particular, stacking regressor – are useful in predicting diabetic nephropathy. The experimental results show that the albumin-to-creatinine (UACR) ratio of urine and urine creatinine works well to detect diabetic nephropathy. The results also analyze the importance of feature selection as a powerful predictor factor. However, based on results, the stacking regressor method does not show a significant improvement compared to Random Forest in terms of predictive outcomes. While the high computation time in training data and inference is associated with accumulation regressors compared to Random Forest, computing efficiency is an important consideration, especially in practical applications where runtime performance requires further research. Therefore, it is required to enhance ensemble learning techniques in predicting diabetic nephropathy, especially in computational time.

6 DATA AVAILABILITY

The dataset and source code of this research can be access at URL: https://github.com/amiragn/Prediction-of-Nephropathy-in-Diabetic-Patients.

7 ACKNOWLEDGMENT

This research is supported by Faculty of Computer Science, Brawijaya University under research grant "Hibah Doktor Lektor Kepala" under contract number 2120/UN10.F15/PN/2023 was dated 6 June, 2023.

8 REFERENCES

[1] H. Sun *et al.*, "IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, p. 109119, 2022. https://doi.org/10.1016/j.diabres.2021.109119

[2] M. K. Sagoo and L. Gnudi, "Diabetic nephropathy: An overview," in *Diabetic Nephropathy: Methods and Protocols*, L. Gnudi and D. A. Long, Eds., in Methods in Molecular Biology, New York, NY: Springer US, 2020, pp. 3–7. https://doi.org/10.1007/978-1-4939-9841-8_1

[3] M. C. Pelle *et al.*, "Up-date on diabetic nephropathy," *Life*, vol. 12, no. 8, p. 1202, 2022. https://doi.org/10.3390/life12081202

[4] N. M. Selby and M. W. Taal, "An updated overview of diabetic nephropathy: Diagnosis, prognosis, treatment goals and latest guidelines," *Diabetes Obes Metab*, vol. 22, no. Suppl 1, pp. 3–15, 2020. https://doi.org/10.1111/dom.14007

[5] N. Samsu, "Diabetic nephropathy: Challenges in pathogenesis, diagnosis, and treatment," *BioMed Research International*, vol. 2021, p. e1497449, 2021. https://doi.org/10.1155/2021/1497449

[6] G. S. Collins and K. G. M. Moons, "Reporting of artificial intelligence prediction models," *The Lancet*, vol. 393, no. 10181, pp. 1577–1579, 2019. https://doi.org/10.1016/S0140-6736(19)30037-6

[7] A. H. Aljammal, S. Taamneh, A. Qawasmeh, and H. B. Salameh, "Machine learning based phishing attacks detection using multiple datasets," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 5, pp. 71–83, 2023. https://doi.org/10.3991/ijim.v17i05.37575

[8] A. Bodor, M. Hnida, and N. Daoudi, "Machine learning models monitoring in MLOps context: Metrics and tools," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 23, pp. 125–139, 2023. https://doi.org/10.3991/ijim.v17i23.43479

[9] A. H. Aljammal, A. Qawasmeh, A. Mughaid, S. Taamneh, F. I. Wedyan, and M. Obiedat, "Performance evaluation of machine learning approaches in detecting IoT-botnet attacks," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 19, pp. 136–146, 2023. https://doi.org/10.3991/ijim.v17i19.41379

[10] Md. Maniruzzaman, M. N. Islam, Md. J. Rahman, Md. A. M. Hasan, and J. Shin, "Risk prediction of diabetic nephropathy using machine learning techniques: A pilot study with secondary data," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 5, p. 102263, 2021. https://doi.org/10.1016/j.dsx.2021.102263

[11] A. Nicolucci *et al.*, "Prediction of complications of type 2 diabetes: A machine learning approach," *Diabetes Research and Clinical Practice*, vol. 190, p. 110013, 2022. https://doi.org/10.1016/j.diabres.2022.110013

[12] S. Ardchir, Y. Ouassit, S. Ounacer, M. Y. E. Ghoumari, and M. Azzouazi, "An integrated ensemble learning framework for predicting liver disease," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 13, pp. 138–152, 2023. https://doi.org/10.3991/ijoe.v19i13.41871

[13] V. MR and J. Chandrika, "EANNMHO – A novel ensemble based technique for liver cirrhosis detection," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 17, no. 4, p. 96, 2021. https://doi.org/10.3991/ijoe.v17i04.21027

[14] H. Kumarasinghe, S. Kolonne, C. Fernando, and D. Meedeniya, "U-net based chest X-ray segmentation with ensemble classification for Covid-19 and Pneumonia," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 7, pp. 161–175, 2022. https://doi.org/10.3991/ijoe.v18i07.30807

[15] K. Gadgil, S. S. Gill, and A. M. Abdelmoniem, "A meta-learning based stacked regression approach for customer lifetime value prediction," *Journal of Economy and Technology*, vol. 1, pp. 197–207, 2023. https://doi.org/10.1016/j.ject.2023.09.001

[16] D. Cui, Z. Wang, P. Liu, S. Wang, Y. Zhao, and W. Zhan, "Stacking regression technology with event profile for electric vehicle fast charging behavior prediction," *Applied Energy*, vol. 336, p. 120798, 2023. https://doi.org/10.1016/j.apenergy.2023.120798

[17] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: An overview," *Journal of Thoracic Disease*, vol. 11, no. Suppl 4, pp. S574–S584, 2019. https://doi.org/10.21037/jtd.2019.01.25

[18] F. Dharma, S. Shabrina, A. Noviana, M. Tahir, N. Hendrastuty, and W. Wahyono, "Prediction of Indonesian inflation rate using regression model based on genetic algorithms," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 45–52, 2020.

[19] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, pp. 1–11, 2020. https://doi.org/10.1155/2020/8885861

[20] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022. https://doi.org/10.1016/j.engappai.2022.105151

[21] Y. Feng and Q. Wu, "A statistical learning assessment of Huber regression," *Journal of Approximation Theory*, vol. 273, p. 105660, 2022. https://doi.org/10.1016/j.jat.2021.105660

[22] C. Dewi and R.-C. Chen, "Random forest and support vector machine on features selection for regression analysis," *International Journal of Innovative Computing, Information and Control*, vol. 15, no. 6, pp. 2027–2037, 2019. https://doi.org/10.24507/ijicic.15.06.2027

[23] H. Yang, Z. Liu, Y. Li, H. Wei, and N. Huang, "CatBoost–bayesian hybrid model adaptively coupled with modified theoretical equations for estimating the undrained shear strength of clay," *Applied Sciences*, vol. 13, no. 9, p. 5418, 2023. https://doi.org/10.3390/app13095418

[24] L. Yao, Z. Fang, Y. Xiao, J. Hou, and Z. Fu, "An Intelligent fault diagnosis method for lithium battery systems based on grid search support vector machine," *Energy*, vol. 214, p. 118866, 2021. https://doi.org/10.1016/j.energy.2020.118866

[25] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 2019, pp. 1–8. https://doi.org/10.1109/ICACCP.2019.8882943

[26] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021. https://doi.org/10.7717/peerj-cs.623

9 AUTHORS

Lailil Muflikhah received a B.Sc. degree in computer science from the Institute Technologi Sepuluh Nopember, an M.Sc. degree in computer science from the Universiti Technology of Petronas, Malaysia, and a Ph.D. degree in Bioinformatics from The University of Brawijaya. She is currently an Associate Professor with the Department of Informatics Engineering in the Faculty of Computer Science,

University of Brawijaya. Her research interests include soft computing, machine learning, and intelligent systems. She can be contacted at lailil@ub.ac.id.

Amira Ghina Nurfansepta is a student of Informatics Engineering at Faculty of Computer Science, Brawijaya University. She has experience as a research assistant. Her research interest is in machine learning, She can be contacted at amiragn25@student.ub.ac.id.

Fitra Abdurrachman Bachtiar is a lecturer in the Faculty of Computer Science at Brawijaya University. He graduated in Electrical Engineering from Brawijaya University in 2008 and from Ritsumeikan University Graduate School of Science and Engineering in 2011. He received the Dr. Eng from Ritsumeikan University in 2016. His research interests are affective computing, affective engineering, intelligent systems, and data mining. He can be contacted at fitra.bachtiar@ub.ac.id.

Dian Eka Ratnawati is an educator at the Faculty of Computer Science, Brawijaya University, Indonesia. She earned her bachelor's and master's degrees from the Sepuluh Nopember Institute of Technology (ITS). Her master's studies focused on Computer Science at ITS, with research centered on Hybrid Genetic Algorithms for job shop scheduling. Continuing her academic journey, Dian pursued her Ph.D. at Brawijaya University (UB), delving into the field of Bioinformatics. Her research specifically explored the utilization of Neural Networks (ELM) and PSO Optimization for drug prediction through SMILES code. She can be contacted at dian_ilkom@ub.ac.id.