ICIC Express Letters
Volume **16**, Number **12**, December **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **1341–1349**

DIABETES PREDICTION USING ENSEMBLE STACKING WITH LASSO AND GENETIC ALGORITHM FOR FEATURE SELECTION

Wiryanata Chandra* and Sani Muhamad Isa

Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
sani.m.isa@binus.ac.id
*Corresponding author: wiryanata.chandra@binus.ac.id

Received December 2021; accepted March 2022

Abstract. *Diabetes mellitus is a metabolic disorder disease that is caused by the pancreas being unable to produce insulin or the body unable to use the insulin that has been produced. Diabetes is dangerous for a person because it can lead to complications of other diseases and even death. The purpose of this study is to create a more accurate diabetes prediction model in order to detect diabetes before other complications occur. Ensemble stacking is proposed to predict diabetes with the help of feature selection, namely Least Absolute Shrinkage and Selection Operator (LASSO) and Genetic Algorithm. For the experiment, National Health and Nutrition Examination Survey (NHANES) is used as the dataset and pre-processed before entering ensemble stacking model. The final result shows that this ensemble stacking model which consists of Naïve Bayes, decision tree and SVM as the base classifier is able to achieve the highest accuracy of 97.97%.*
Keywords: Diabetes prediction, LASSO, Genetic Algorithm, Ensemble stacking

1. **Introduction.** Diabetes mellitus is a metabolic disorder disease caused by the pancreas that cannot produce enough insulin or the body cannot use the insulin produced efficiently and effectively, resulting in an increase in glucose levels in the blood. This disease is a chronic disease and the number of patients worldwide is quite large [1]. Diabetes mellitus is divided into 2 types. Type 1 occurs when the body fails to produce insulin completely. So this is often referred to as "*Insulin Dependent Diabetes Mellitus*". Type 2 occurs when the body cannot use the insulin it produces effectively. This type is referred to as "*Non-Insulin Dependent Diabetes Mellitus*". And the third type is diabetes which occurs when a pregnant woman receives insulin [2]. Of all people with diabetes, 95% of them are people with type 2 diabetes, where this percentage will increase every year and is expected to double in the next 15 years [3]. Diabetes is a dangerous disease because it can lead to complications of other diseases and even death [4]. So the need here is that we are able to detect or predict this diabetes before other complications occur.

Studies on the prediction of diabetes have been done quite a lot by applying several algorithms of machine learning. There are even some studies that use the ensemble model. In a study conducted by [5], they predict diabetes using the TPASVM (Tree Partitioning Adaptive Support Vector Machine) algorithm. By using the SMORT technique in partitioning the tree, they improved the accuracy around 8.67%. Likewise, what was done by [6] is that they used the ensemble technique to predict type 1 diabetes. The results they got were that using the bagging metaregressor method could increase the accuracy by 4%-6%.

DOI: 10.24507/icicel.16.12.1341

So the main contribution in this study is trying to evaluate an ensemble stacking model using the NHANES dataset used in [7]. The ensemble stacking model proposed in this study is to combine Naïve Bayes, decision tree, and SVM as the base classifier. And later the meta classifier will be compared between the three models and decide which is the best result as the proposed model. In addition, this study will also use several feature selection methods that were not used in previous studies, that are LASSO (Least Absolute Shrinkage and Selection Operator) and Genetic Algorithm in order to increase accuracy of the proposed model.

This paper is organized as follows. Section 2 reviews the literature in the domain of diabetes prediction. In Section 3, we introduce the proposed methodology. Then Section 4 discusses results of the experiment. In the end, the conclusion and future work are presented in Section 5.

2. **Related Works.** A lot of research on diabetes prediction has been done, especially using machine learning methods. The implementation of machine learning methods, especially in predicting diabetes, often uses the Pima Indian Diabetes Dataset [8] or datasets such as questionnaires [9]. Several studies have implemented simple machine learning models such as KNN (K-Nearest Neighbor) which outperform several models such as binary logistic regression and multilayer perceptron with an accuracy of 80% [10]. There is also a study that used improved Naïve Bayes [11] and SVM [29] which achieves an accuracy of 82.30% for improved Naïve Bayes and 76.30% for SVM. In addition, when implementing this machine learning model, problems that often occur are the number of data outliers, imbalances in the dataset and also the many dimensions of the dataset used. IQR (Interquartile Range) and SMOTE (Synthetic Minority Oversampling Technique) is one solution to prevent this problem [12] which achieves an accuracy of 89.50% with the C4.5 model. In addition, PCA (Principal Component Analysis) is also used as a solution to reduce the dimensionality of the dataset used [13,14].

In addition to research that uses simple machine learning methods, there are also studies that use the ensemble method. Ensemble methods such as the soft voting achieve 79.08% accuracy [15]. In addition, there is also a study that proposes an ensemble stacking model that uses AdaBoost as a meta classifier which achieves 90.36% of accuracy [16]. There are also studies that propose a model called EasyEnsemble [17] and also XGBoost [18]. Just like the single model, this ensemble method also has several problems, such as too many dataset dimensions and many unused features. Feature selection methods such as Chi Square [19] and weighted feature selection [20] are problem solvers, especially for unused features. In addition, just like the single model, PCA and mRMR (Minimum Redundancy Maximum Relevance) are used to reduce the dimensionality of the dataset which achieves an accuracy of 77.21% [21].

Based on all these studies, we know that ensemble technique will significantly increase the accuracy of the model especially in this diabetes prediction case. So, in this study we will be using an ensemble technique that is called ensemble stacking and combining it with feature selection methods such as LASSO and Genetic Algorithm on NHANES Dataset (2017-2018).

3. **Methodology.** The experimental process that is proposed in this study can be seen in Figure 1. This experiment started with data collection, data labeling, data pre-processing for the three schemes, feature selection (LASSO and Genetic Algorithm), data splitting, training and evaluating single models, and finally training and evaluating ensemble stacking models.

3.1. **Data collection.** The dataset used in this study was obtained from a survey conducted to determine the health condition and also the amount of nutrition possessed by

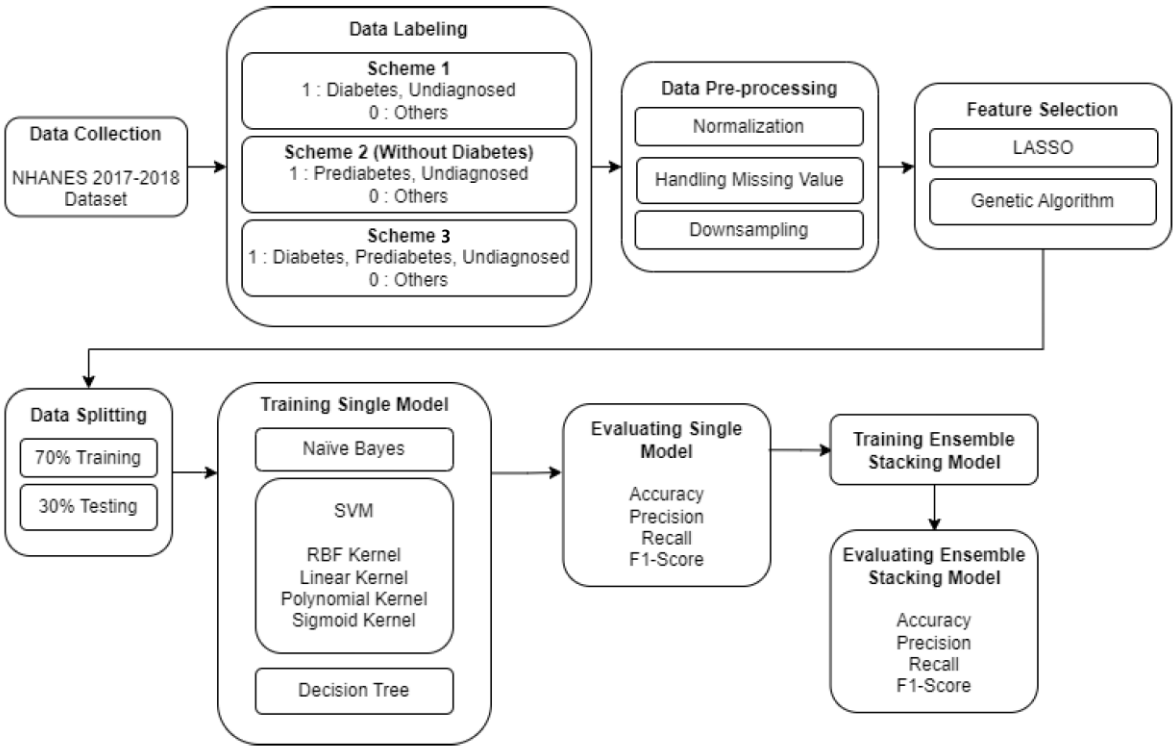

FIGURE 1. Proposed methodology

children to adults in the United States that is called NHANES (National Health and Nutrition Examination Survey) dataset. NHANES is a cross-sectional survey that aims at evaluating the health and nutritional status of Americans [22]. This survey is quite unique because it combines an interview and a physical examination of a person. The interview data was examined from a subset comprising adults (age $\geq$ 18 years) [23]. In this study, the NHANES data used was in the 2017-2018 and only the laboratory data and questionnaire data will be used. For the number of available attributes when combined between laboratory data and questionnaire, there are 56 attributes with 8897 total records.

3.2. **Data labeling.** The data labeling process is carried out with reference to [7]. In their research, they divided the datasets into three schemas. The first was to classify the diabetic subjects and the second was to classify the pre-diabetic/undiagnosed diabetic subjects. In the first scheme, the subject is categorized as diabetic (label = 1) if it meets the requirements, such as answering "*Yes*" to the question "*Have you ever been told by a doctor that you have diabetes?*" or have a blood glucose level of more than or equal to 126 mg/dL; otherwise it will be categorized as non-diabetic subjects (label = 0). In this first scheme, there are 1023 diabetic subjects (label = 1) and 7874 non-diabetic subjects (label = 0). Then on the second scheme, the undiagnosed category (label = 1) is given if the subject answered "*No*" to the question "*Have you ever been told by a doctor that you have diabetes?*" and have a blood glucose level of more than or equal to 126 mg/dL, then the pre-diabetes category (label = 1) if the blood glucose level is between 100 and 125 mg/dL. Specifically for this scheme, all subjects who have been diagnosed with diabetes will not be included. Apart from this, the subject was categorized as non-diabetic (label = 0). In this second scheme, there are 1738 pre-diabetic/undiagnosed subjects (label = 1) and 6266 non-diabetes subjects (label = 0). Apart from the division of the scheme as done by [7], this study will also add 1 more scheme, to classify subjects with diabetes and pre-diabetic/undiagnosed. This third scheme is like a combination of the two schemes that have been made previously. So there are 2631 diabetic/pre-diabetic/undiagnosed

TABLE 1. Label assignments

Classification	Scheme 1	Scheme 2	Scheme 3
Diabetic	1	Excluded	1
Pre-diabetic	0	1	1
Non-diabetic	0	0	0
Undiagnosed	0	1	1

(label = 1) subjects and 6266 non-diabetic subjects (label = 0). The corresponding label assignments for each case are shown in Table 1.

3.3. **Data pre-processing.** After getting the dataset, the next step is to process the data to suit our needs. In this case, there are several stages to be carried out, including handling missing values, normalization and also down sampling.

This missing value handling is done because in this dataset there are a lot of missing values on several attributes because basically this dataset is a questionnaire so it is very possible for many missing values to occur. To solve this missing value, an imputation process will be carried out which uses the mean/mode of each attribute. For some cases such as questions that can only be answered with "*Yes No Question*" it is not possible to use the mean, so the mode of the attribute will be used. Likewise, for the other questions the mean of the attribute will be used to solve this missing value.

Then after handling the missing value, the data normalization process will be carried out. The main goal of this process is to limit the value of an attribute to be within a certain range. Because in this dataset there are several attributes that have values with a wide enough range so it will reduce the performance of the model that will be created. Normalization is done by using a defined range between 1-10.

Then the last one is down sampling where this is done because the number of labels on the data has a quite difference. The down sampling technique used in this research is bootstrap down sampling. This is done because this method is quite simple and easy to implement.

3.4. **Feature selection.** The next process to be carried out is to select features/attributes that have a high correlation with the labels that have been previously created and then the other features will be removed [24]. In this study, there are 2 methods that will be used, i.e., LASSO and Genetic Algorithm.

LASSO (Least Absolute Selection and Shrinkage Operator) is a method used to select correlated features by combining 2 feature selection concepts, namely filter methods and wrapper methods. This combined concept is referred to as embedded methods [25]. LASSO is included in one part of linear regression where the difference is in the penalty term. The penalty term in LASSO is what we can optimize in order to achieve maximum performance. Parameter selection here is done using the GridSearchCV with total of cross validation of 10 which looks for the best parameters between 0.00001 to 0.9. Based on the results, so in this study the penalty term to be used is $\lambda = 0.00001$.

Genetic Algorithm is one of the algorithms or methods to perform a heuristic search which is used to search for the most optimal solution [26]. There are several important processes in this Genetic Algorithm that are initial population, fitness function calculation, selection, crossover and mutation. And also there are few important parameters in this Genetic Algorithm such as total generation, total population, crossover rate and also mutation rate. For the selection of the most optimal parameters, we also use GridSearchCV with total cross validation of 10 for changing the number of population and generation only. As for the mutation and crossover rate use the default values by the library, namely 0.9 and 0.1. So based on the results, parameter that will be used in this study are 30 generation, 30 population, 0.9 crossover rate and 0.1 mutation rate.

3.5. **Data splitting.** After the best features have been selected in the previous process, then the next step is to separate the training data and also the testing data. The proportion of training data is 70% and for testing data 30%.

3.6. **Training and evaluation.** Then the last stage is the training process and also the evaluation of the model, either single model or ensemble stacking model. The first process that will be carried out is training on a single model, namely decision tree, Naïve Bayes and support vector machine (kernel = *linear*, *polynomial*, *RBF*, *sigmoid*). After the training process is carried out, the results will be evaluated using testing data that has been split previously by considering 4 performance metrics, namely *accuracy*, *precision*, *recall* and *f1-score*.

Finally, single models will be combined to create an ensemble model, namely stacking. Ensemble model is one of the classification techniques in which multiple classification algorithms will be combined in order to reduce bias and variance and to improve performance in terms of accuracy [27]. Ensemble stacking model will combine the training results from several single models (*base classifier*) and then make them as input to other models (*meta classifier*) [28]. In this study, there are 3 base classifiers used, i.e., decision tree, Naïve Bayes and support vector machine. Then the results of the ensemble stacking model will be evaluated the same as the previous single model. The ensemble stacking architecture can be seen in Figure 2 below.

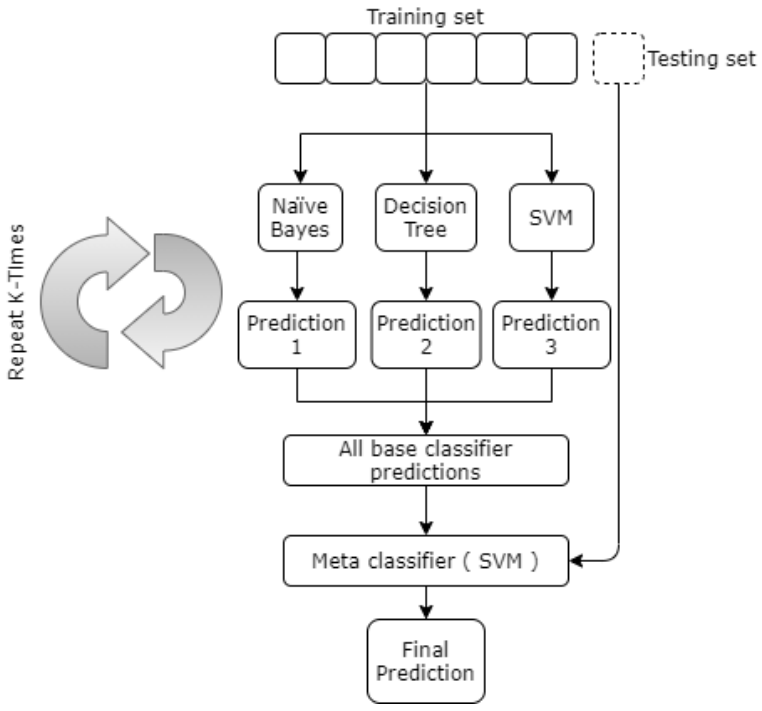

FIGURE 2. Ensemble stacking architecture

4. **Results and Discussion.** The first evaluation was carried out on a single model using the Naïve Bayes, SVM and also the decision tree algorithm. Especially for the SVM algorithm, experiments were carried out using several provided kernels such as RBF, Linear, Polynomial and Sigmoid. In addition, experiments using this single model were carried out on the 3 schemes described earlier. The results of the evaluation can be seen in Table 2.

Based on Table 2, we can see that in scheme 1 the single model has an accuracy that is not too significant, which is around 94.01% to 94.40% except for the SVM model with a sigmoid kernel which has an accuracy of only 60.87%. Then in scheme 2, the single

Table 2. Single model evaluation results

Scheme	Model	Accuracy	Precision	Recall	F1-score
1	Naïve Bayes	94.01%	99.81%	88.19%	93.64%
	Decision tree	94.40%	98.66%	90.64%	94.48%
	SVM (RBF kernel)	94.03%	99.86%	88.19%	93.66%
	SVM (Linear kernel)	94.31%	99.25%	89.29%	94.00%
	SVM (Poly kernel)	94.10%	99.90%	88.27%	93.73%
	SVM (Sigmoid kernel)	60.87%	62.12%	55.67%	58.72%
2	Naïve Bayes	52.66%	51.62%	94.32%	66.72%
	Decision tree	68.80%	71.16%	72.90%	72.02%
	SVM (RBF kernel)	68.52%	62.23%	95.26%	75.28%
	SVM (Linear kernel)	68.14%	62.39%	92.34%	74.47%
	SVM (Poly kernel)	68.75%	62.21%	96.57%	75.67%
	SVM (Sigmoid kernel)	64.36%	67.43%	56.41%	61.43%
3	Naïve Bayes	66.62%	97.71%	33.87%	50.59%
	Decision tree	77.39%	56.32%	58.85%	73.06%
	SVM (RBF kernel)	77.98%	55.36%	58.17%	72.26%
	SVM (Linear kernel)	78.03%	64.99%	58.89%	72.47%
	SVM (Poly kernel)	78.00%	65.38%	58.38%	72.43%
	SVM (Sigmoid kernel)	66.76%	58.91%	39.90%	55.08%

model is able to predict with the greatest accuracy achieved by the decision tree model of 68.80%. Meanwhile, for the highest f1-score value, the SVM model using a polynomial kernel is obtained, which is 75.67%. While in the last scheme, namely scheme 3, the single model achieves the highest accuracy and also f1-score by using the SVM model with a linear kernel which are 78.03% and 72.47%, respectively. However, there is a significant difference between precision and recall in the Naïve Bayes classifier in scheme 3, which is about 60% difference. This can happen because this Naïve Bayes classifier gives a higher false negative value so that is why the precision can be higher by around 60% compared to recall. In these three schemes, the single model is able to achieve the highest accuracy in scheme 1, namely the decision tree model of 94.40%. This can happen because of the influence of the dataset which was originally distributed better in scheme 1 so that this makes the accuracy of the model indeed much higher than in other schemes.

Then the next evaluation is carried out using one of the ensemble models which is specifically designed to increase the accuracy of a model. The ensemble model is the stacking model. This ensemble stacking model consists of 2 levels of classifier, the base classifier and the meta classifier. The base classifier used for this ensemble stacking model is obtained from all single models that have been evaluated previously. And SVM with linear kernel will be used as a meta classifier on ensemble stacking model. However, specifically for the base classifier, the SVM model used will only use a linear kernel. This is because Table 2 shows that the greatest potential of SVM is in the linear kernel by achieving a fairly high accuracy compared to other kernels. After evaluating the ensemble stacking model, the ensemble stacking model will then be added with the feature selection method. Feature selection that will be used in this evaluation is LASSO and also Genetic Algorithm. Because the main goal of this stacking ensemble is to increase the accuracy of the prediction model, Table 3 will show a comparison of each model in several schemes.

Based on the table, it can be seen that in the first scheme, the greatest accuracy of 97.97% is achieved by the ensemble stacking model plus LASSO. Then followed by the ensemble stacking model plus Genetic Algorithm which is 96.84%. In terms of precision, the best performance was achieved by the ensemble stacking model of 99.90% and the lowest performance was obtained by the ensemble stacking plus the Genetic Algorithm

Table 3. Final evaluation results

Scheme	Model	Accuracy	Precision	Recall	F1-score
1	Naïve Bayes	94.01%	99.81%	88.19%	93.64%
	Decision tree	94.40%	98.86%	90.64%	94.48%
	SVM (Linear kernel)	94.31%	99.90%	89.29%	94.00%
	Ensemble stacking	94.73%	99.90%	94.33%	94.40%
	Ensemble stacking + LASSO	97.97%	97.94%	97.23%	97.95%
	Ensemble stacking + GA	96.84%	96.67%	96.71%	96.84%
2	Naïve Bayes	52.66%	51.62%	94.32%	66.72%
	Decision tree	68.80%	71.16%	72.90%	72.02%
	SVM (Linear kernel)	68.75%	67.43%	96.57%	75.67%
	Ensemble stacking	71.07%	72.62%	99.76%	76.25%
	Ensemble stacking + LASSO	96.63%	95.58%	98.42%	96.20%
	Ensemble stacking + GA	88.22%	86.92%	88.99%	87.94%
3	Naïve Bayes	66.62%	97.71%	33.87%	50.59%
	Decision tree	77.39%	56.32%	58.85%	73.06%
	SVM (Linear kernel)	78.03%	64.99%	58.89%	72.47%
	Ensemble stacking	79.47%	93.66%	92.55%	76.75%
	Ensemble stacking + LASSO	97.62%	96.44%	96.32%	97.20%
	Ensemble stacking + GA	93.22%	93.63%	92.54%	93.08%

which was 96.67%. In terms of recall, the highest performance is on the ensemble stacking model plus LASSO of 97.23% and the lowest is on the Naïve Bayes model, which is 88.19%. Then the largest f1-score value is in ensemble stacking added with LASSO of 97.95% and the lowest is in Naïve Bayes which is 93.64%. In the second scheme, the greatest accuracy is in the ensemble stacking model added with LASSO which is 96.63% and the smallest accuracy is in the Naïve Bayes model of 52.66%. Then the highest precision is achieved by the ensemble stacking model plus LASSO which is 95.58%, followed by ensemble stacking plus the Genetic Algorithm which is 86.92%. As for recall and also the highest f1-score it is achieved by the ensemble stacking model plus LASSO at 98.42% and 96.20%, respectively. In the third scheme, which is the last one, it is dominated by the ensemble stacking model of its performance. All aspects such as accuracy, recall and also f1-score except for precision outperformed by Naïve Bayes, which is 97.71%. And also, in terms of the complexity of the algorithm itself, ensemble stacking clearly has a higher complexity than simple machine learning algorithms. This is because this stacking ensemble has an architecture that will combine the classification results from several algorithms and be used as input for other algorithms.

5. **Conclusions.** This study illustrates how the ensemble stacking model is able to improve performance both in terms of accuracy and from other aspects of the model to predict diabetes when compared to a single model. In addition to this ensemble stacking model, there is an important role of the feature selection method as well to improve the performance of the model. This study proposes a method using ensemble stacking in which several single models are used, namely Naïve Bayes, support vector machine and decision tree which are then added with feature selection methods, namely LASSO and Genetic Algorithm. This stacking ensemble is able to increase the accuracy with an average of 0.66% when compared to the single model. However, if added with feature selection LASSO or Genetic Algorithm, the accuracy of the ensemble stacking model is able to achieve an accuracy of around 97% in the three schemes.

For future work, there is still a lot that can be explored further in this research, such as the type of ensemble model used, the combination of single models in it, and especially the

use of larger and up-to-date datasets that can be a challenge in future research because of the limitations of the datasets used in this study.

REFERENCES

[1] S. Perveen, M. Shahbaz, A. Guergachi and K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, *Procedia Computer Science*, vol.82, pp.115-121, DOI: 10.1016/j.procs.2016.04.016, 2016.

[2] C. Kalaiselvi and G. M. Nasira, A new approach for diagnosis of diabetes and prediction of cancer using ANFIS, *2014 World Congress on Computing and Communication Technologies*, Trichirappalli, India, pp.188-190, DOI: 10.1109/WCCCT.2014.66, 2014.

[3] B. V. Baiju and D. J. Aravindhar, Disease influence measure based diabetic prediction with medical data set using data mining, *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, Chennai, India, pp.1-6, DOI: 10.1109/ICIICT1.2019.8741452, 2019.

[4] M. S. Kadhm, An accurate diabetes prediction system based on K-means clustering and proposed classification approach, *International Journal of Applied Engineering Research*, vol.13, no.6, pp.4038-4041, 2018.

[5] R. Syed, R. K. Gupta and N. Pathik, An advance tree adaptive data classification for the diabetes disease prediction, *2018 International Conference on Recent Innovations in Electrical, Electronics and Communication Engineering (ICRIEECE)*, Bhubaneswar, India, pp.1793-1798, DOI: 10.1109/ICRIEECE44171.2018.9009180, 2018.

[6] K. Saiti, M. Macaš, L. Lhotská, K. Štechová and P. Pithová, Ensemble methods in combination with compartment models for blood glucose level prediction in type 1 diabetes mellitus, *Computer Methods and Programs in Biomedicine*, vol.196, 105628, DOI: 10.1016/j.cmpb.2020.105628, 2020.

[7] A. Dinh, S. Miertschin, A. Young and S. D. Mohanty, A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, *BMC Medical Informatics and Decision Making*, vol.19, no.1, pp.1-15, DOI: 10.1186/s12911-019-0918-5, 2019.

[8] R. N. Kumar and M. A. Kumar, A novel feature selection algorithm with dempster shafer fusion information for medical datasets, *International Journal of Applied Engineering Research*, vol.12, no.14, pp.4205-4212, 2017.

[9] H. Wu, S. Yang, Z. Huang, J. He and X. Wang, Type 2 diabetes mellitus prediction model based on data mining, *Informatics in Medicine Unlocked*, vol.10, pp.100-107, DOI: 10.1016/j.imu.2017.12.006, 2018.

[10] S. Selvakumar, K. S. Kannan and S. Gothainachiyar, Prediction of diabetes diagnosis using classification based data mining techniques, *International Journal of Statistics and Systems*, vol.12, no.2, pp.183-188, http://www.ripublication.com, 2017.

[11] N. Sneha and T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, *Journal of Big Data*, vol.6, no.1, pp.1-19, DOI: 10.1186/s40537-019-0175-6, 2019.

[12] N. Nnamoko and I. Korkontzelos, Efficient treatment of outliers and class imbalance for diabetes prediction, *Artificial Intelligence in Medicine*, vol.104, 101815, DOI: 10.1016/j.artmed.2020.101815, 2020.

[13] G. Geetha and K. M. Prasad, Prediction of diabetes using machine learning, *International Journal of Recent Technology and Engineering*, vol.8, no.5, pp.1119-1124, DOI: 10.35940/ijrte.e6290.018520, 2020.

[14] M. T. Islam, M. Raihan, S. R. I. Akash, F. Farzana and N. Aktar, Diabetes mellitus prediction using ensemble machine learning techniques, *Communications in Computer and Information Science*, vol.1192, pp.453-467, DOI: 10.1007/978-981-15-3666-3_37, 2020.

[15] S. Kumari, D. Kumar and M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, *International Journal of Cognitive Computing in Engineering*, vol.2, pp.40-46, DOI: 10.1016/j.ijcce.2021.01.001, 2021.

[16] M. Alehegn, R. Joshi and P. Mulay, Analysis and prediction of diabetes mellitus using machine learning algorithm, *International Journal of Pure and Applied Mathematics*, vol.118, no.9, pp.871-878, 2018.

[17] T. Yang et al., Ensemble learning models based on noninvasive features for type 2 diabetes screening: Model development and validation, *JMIR Medical Informatics*, vol.8, no.6, pp.1-11, DOI: 10.2196/15431, 2020.

[18] L. Wang, X. Wang, A. Chen, X. Jin and H. Che, Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model, *Healthcare*, vol.8, no.3, pp.1-11, DOI: 10.3390/healthcare8030247, 2020.

[19] T. E. Dagogo-George, H. A. Mojeed, A. O. Balogun, M. A. Mabayoje and S. A. Salihu, Tree-based homogeneous ensemble model with feature selection for diabetic retinopathy prediction, *Jurnal Teknologi dan Sistem Komputer*, vol.8, no.4, pp.297-303, DOI: 10.14710/jtsiskom.2020.13669, 2020.

[20] Z. Xu and Z. Wang, A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and XGBoost ensemble classifier, *2019 11th International Conference on Advanced Computational Intelligence (ICACI)*, Guilin, China, pp.278-283, DOI: 10.1109/ICACI.2019.8778622, 2019.

[21] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, Predicting diabetes mellitus with machine learning techniques, *Frontiers in Genetics*, vol.9, pp.1-10, DOI: 10.3389/fgene.2018.00515, 2018.

[22] P. Fong and Q. T. Wang, Protective effect of oral contraceptive against Helicobacter pylori infection in US adult females: NHANES 1999-2000, *Epidemiology and Infection*, vol.149, e120, DOI: 10.1017/S0950268821000923, 2021.

[23] A. R. Williams, M. Wilson-Genderson and M. D. Thomson, A cross-sectional analysis of associations between lifestyle advice and behavior changes in patients with hypertension or diabetes: NHANES 2015-2018, *Preventive Medicine*, vol.145, 106426, DOI: 10.1016/j.ypmed.2021.106426, 2021.

[24] A. Negi and V. Jaiswal, A first attempt to develop a diabetes prediction method based on different global datasets, *2016 4th International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, India, pp.237-241, DOI: 10.1109/PDGC.2016.7913152, 2016.

[25] V. Fonti and E. Belitser, Feature selection using LASSO, *VU Amsterdam Research Paper in Business Analytics*, vol.30, pp.1-25, 2017.

[26] N. Maleki, Y. Zeinali and S. T. A. Niaki, A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection, *Expert Systems with Applications*, vol.164, 113981, DOI: 10.1016/j.eswa.2020.113981, 2021.

[27] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, Development of disease prediction model based on ensemble learning approach for diabetes and hypertension, *IEEE Access*, vol.7, pp.144777-144789, DOI: 10.1109/ACCESS.2019.2945129, 2019.

[28] S. Agarwal and C. R. Chowdary, A-Stacking and A-Bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection, *Expert Systems with Applications*, vol.146, 113160, DOI: 10.1016/j.eswa.2019.113160, 2020.

[29] A. Chowanda, D. Suhartono, E. W. Andangsari and K. Z. bin Zamli, Machine learning algorithms exploration for predicting personality from text, *ICIC Express Letters*, vol.16, no.2, pp.117-125, DOI: 10.24507/icicel.16.02.117, 2022.