

Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning

Muhammad Munsarif, Muhammad Sam'an, Safuan

Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia

Article Info

Article history:

Received Apr 7, 2022

Revised Jul 28, 2022

Accepted Aug 30, 2022

Keywords:

Credit risk

Embedded technique

Feature selection

Peer to peer lending

Stacking ensemble model

ABSTRACT

Peer to peer lending is famous for easy and fast loans from complicated traditional lending institutions. Therefore, big data and machine learning are needed for credit risk analysis, especially for potential defaulters. However, data imbalance and high computation have a terrible effect on machine learning prediction performance. This paper proposes a stacking ensemble learning with features selection based on embedded techniques (gradient boosted trees (GBDT), random forest (RF), adaptive boosting (AdaBoost), extra gradient boosting (XGBoost), light gradient boosting machine (LGBM), and decision tree (DT)) to predict the credit risk of individual borrowers on peer to peer (P2P) lending. The stacking ensemble model is created from a stack of meta-learners used in feature selection. The feature selection+ stacking model produces an average of 94.54% accuracy and 69.10 s execution time. RF meta-learner+Stacking ensemble is the best classification model, and the LGBM meta-learner+stacking ensemble is the fastest execution time. Based on experimental results, this paper showed that the credit risk prediction for P2P lending could be improved using the stacking ensemble model in addition to proper feature selection.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Muhammad Munsarif

Departement of Informatics, Universitas Muhammadiyah Semarang

Semarang, Indonesia

Email: m.munsarif@unimus.ac.id

1. INTRODUCTION

Peer to peer (P2P) lending platform as a modern banking system launched to overcome the complexity of conventional loans with the concept of borrowing and borrowing money directly without intermediaries. The growth of P2P has been very rapid since Lehman introduced it in 2008-2009. The complexity of loan transactions is the reason for lending more to P2P. Moreover, traditional financial institutions do not fully cater to risk-seeking lenders and high-risk borrowers [1]. Therefore, P2P lending platforms are becoming increasingly popular where lenders have more flexibility to pick and choose the desired risk portfolio [2].

Traditional lending institutions play at a low-risk level. Therefore, P2P lending is here to bridge this problem by offering easy lending for small businesses or beginners. Low-interest rates and transaction flexibility are the main attractions of P2P lending. Despite this attraction, P2P lending has not been able to ensure that borrowers are in a hurry to be given a loan. Generally, financial institutions use scorecards that contain statistical information on prospective borrowers for credit risk analysis [3], [4]. Recently, the application of machine learning carried out to predict the credit risk of online loans with a very encouraging performance [5]–[9]. Moreover, the number of features and balanced data (the number of borrowers paid on

time is almost the same as the number of default borrowers) affects the prediction results so that online lending institutions and borrowers can use this opportunity.

Most of the available P2P lending data is unbalanced (some people lost to follow-up, with most borrowers returning on time). If this data is used in machine learning training, the borrower will be classified as a good or no-risk borrower. Of course, machine learning prediction accuracy is very high. However, high accuracy does not guarantee the exact model used. Unbalanced data has the potential to make a habit of predicting default or non-default borrowers [10], [11]. In such circumstances, the model is often fused with over-trained models biased to the dominant classes of the available data. Therefore, how to achieve accurate predictions for bad borrowers is very important. In addition, compared to traditional banking systems, P2P lending does not have sufficient information about financial statistics and historical customer data. In addition, the model should be computationally lightweight. Therefore, finding the essential features to reduce the cost of computing becomes a more pressing issue. Fewer features improve classification accuracy and generalization if appropriately chosen [12], [13].

The imbalance of multi-class data types (the number of samples from one or several classes is greater than the other) becomes a challenge for the prediction process. This data imbalance can potentially reduce the model's prediction performance [14]. Therefore, several studies carried out the pre-processing data stage to make the data balanced [15]. The approach used to reduce the dimensions of P2P lending data is by selecting features [16]. The working concept of this approach is to choose features that are considered important and remove features that are not important in the prediction process. Removing non-essential attributes has several advantages, such as reducing memory and computational costs, taking full advantage of precision, and staying clear of over-fitting problems during the training stage [17]. On the other hand, some features might serve for algorithms (e.g., decision tree (DT)). May not be practical for various other models, such as regression models. In addition, irrelevant features can negatively affect model performance. Therefore, data pre-processing and feature selection are the most significant steps in designing and selecting the best model for a particular problem [18].

Well-predictive performance can be achieved through a feature selection approach [19]. Wrapper, embedded, shuffle, and hybrid are types of this approach. The main objective of this research is to improve the model's performance, avoid over-fitting problems, and reduce the dimensions of the input data. Although feature selection has certain drawbacks, it is an important pre-processing technique for ML. It generates additional information and provides an intuitive understanding of typical patterns before the proposed classifier is used [20].

This study uses the embedded technique, i.e., random forest (RF) importance [21]–[25] and boosting (gradient boosted trees (GBDT), extra trees (ET), adaptive boosting (AdaBoost), extra gradient boosting (XGBoost), light gradient boosting machine (LGBM), and decision tree (DT)) importance [26]–[30] for feature selection to improve credit risk identification in P2P lending. This study proposes a Stacking ensemble approach from several machine learning techniques: GBDT, RF, AdaBoost, XGBoost, LGBM, and DT compares their performance. The most commonly used matrix is accuracy to assess the proposed implementation. This paper proposes two contributions: i) selection of features based on embedding technique and stacking ensemble learning model as a classifier and ii) feature selection using embedding technique.

2. METHOD

The P2P solid classification-based ML ensemble model is described in this section. In general, the proposed model is used for the training and testing process of the collected data. Meanwhile, k-fold cross-validation (CV) is used to overcome overfitting in training by setting the average performance classifier. The basic idea of this tool is iteratively repeating 3 times and testing on the fifth iteration. In this study, we apply feature selection to select essential features in credit risk prediction to improve prediction performance. For feature selection, we tested embedded techniques involving GBDT, RF, AdaBoost, XGBoost, LGBM, and DT. The workflow of the proposed framework is presented in Figure 1.

2.1. Data collection and pre-processing

The original data set was collected from the Lending Club website, one of the most popular P2P platforms in the US. The raw data period is from the first quarter of 2019 to the fourth quarter of 2019, containing 42,538 borrowers with 161 features. Initial exploration of the dataset revealed many columns having missing values of more than 68.51%, thus removed. As a result, the dataset's features reduce from 127, the number of features to 34. The predictive model may become too complex. Thus, references to the most recent literature are referenced to delimit the feature space further.

2.2. Feature selection

Personal component analysis (PCA) is used to reduce the dimensions of the data in feature selection. The low-dimensional features that are mapped have no significant effect. PCA cannot distinguish the importance of features in the classification process. Especially in deciding to give or not get a loan, so using essential feature selection techniques is needed.

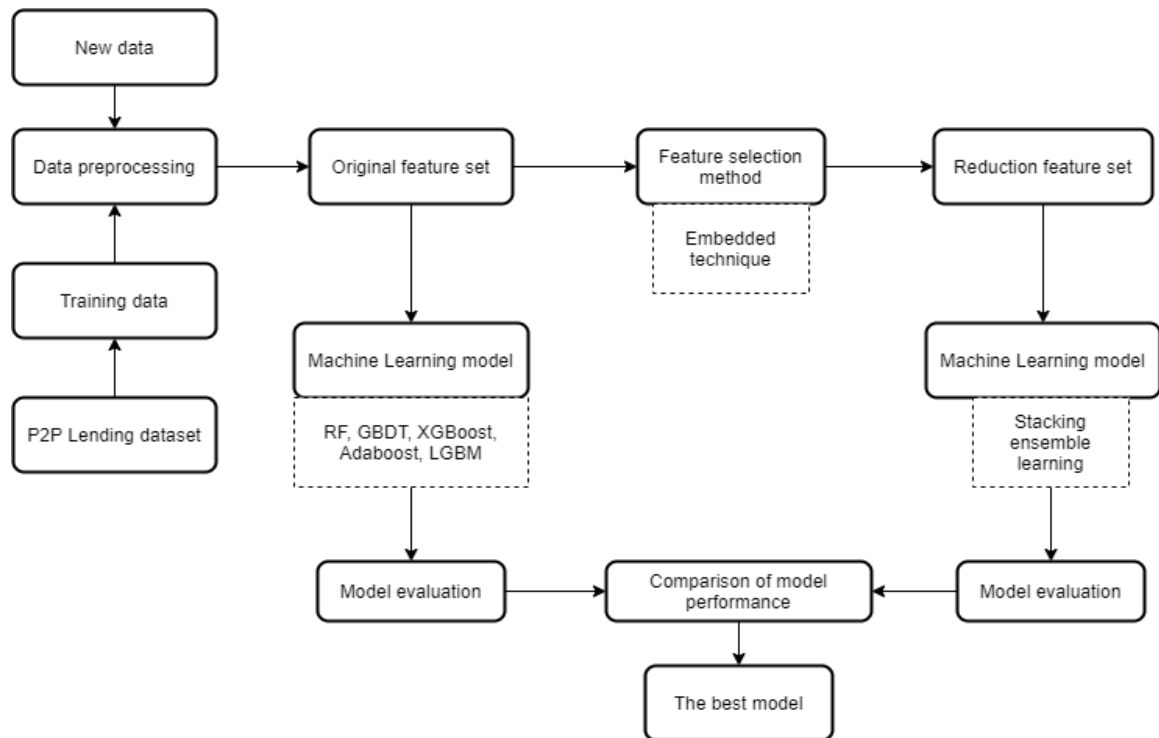


Figure 1. Process steps for implementing feature selection methods and stacking ensemble learning model

Another approach is feature selection. One approach to feature selection is the embedded technique. Embedded techniques complete the machine learning algorithm construction's feature selection process. In other words, they perform feature selection during model training, which is why we call them embedded methods. A learning algorithm takes advantage of its variable selection process and simultaneously performs feature selection and classification or regression. All embedded methods work: first, they train machine learning models. They then derived the important features of this model, which measures how important the features are when making predictions. Finally, they remove unimportant features using important child features.

2.3. Stacking ensemble model

Furthermore Martin *et al.* [31] introduced the stacking method as an ensemble algorithm distinct from bagging, RF, and boosting: stacking considers heterogeneous learners. The schematic diagram of the stacking method is shown in Figure 2. There are usually two or more levels of the classifier. The first level is zero and contains basic classifiers that take the original input. As seen in Figure 2, H_0 is the original dataset, which is the P2P lending dataset in our problem. The zero-level classifier will generate the H_1 dataset, which will be used in the second level by the meta classifier (or first-level classifier). H_1 is the dataset generated by the base classifiers: GBDT, RF, AdaBoost, XGBoost, LGBM, and DT. CB_i is the primary classifier will be used to generate the H_1 dataset. CM_i is the meta-classifier that will be used to classify the H_1 dataset. H_1 can be a probability or a label, meaning the output from CB_i that CM_i will use. We will compare the two methods.

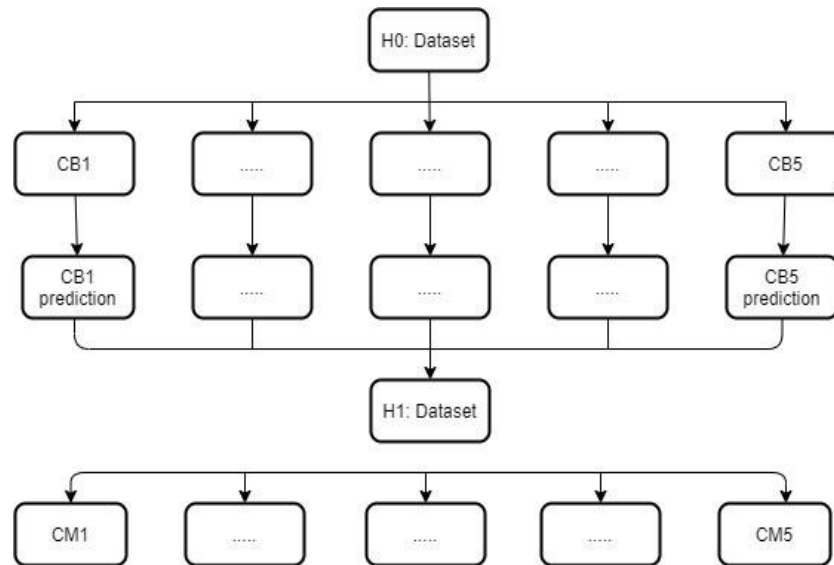


Figure 2. A schematic diagram of the stacking ensemble model

3. RESULTS AND DISCUSSION

The experimental results will be discussed in the following section. All experiments were carried out using a known P2P lending binary dataset. In order to measure the performance of the ML model, 3-fold cross-validation (3-fold CV) was used to calculate mean accuracy. Determining the best features uses the feature importance algorithm from embedded techniques involving GBDT, RF, AdaBoost, XGBoost, LGBM, and DT. Important features are chosen based on the weight value of each feature generated during the predictive analysis process. Details of the best feature selection results are shown in Figure 3 (in appendix).

Figure 3 shows that each meta-learner in the embedded technique produces various important features based on the weight value of each feature so that the use of meta-learners dramatically affects the accuracy score generated by the evaluation model. This study used a stacking ensemble learning model created from a stack of meta-learners used in feature selection. The results of comparing the accuracy scores and execution time of the stacking ensemble model based on the type of meta-learners used in the feature selection process shown in Table 1.

Based on Table 1, the stacking ensemble model produces an average of 94.54% accuracy and 69.10 execution time. RF meta-learner+stacking ensemble is the best classification model, and the LGBM meta-learner+stacking ensemble has the fastest execution time. Meanwhile, when compared to the prediction accuracy of the feature selection+stacking model with the original model, the feature selection+stacking model succeeded in increasing the accuracy of the original model with an average difference from the accuracy of the original model to the feature selection model+stacking model reaching 1.22%. DT is the best meta-learner for increased accuracy on the original model.

Furthermore, the feature selection+stacking model is not efficient on execution time. The original model requires a more efficient execution time than the feature selection+stacking model. However, these limitations do not significantly affect feature selection because the time difference between the original model and the feature selection+model stacking is not too far away and is still acceptable in the computational process. In detail, the comparison of the feature selection+stacking model and the original model can be seen in Table 1.

Table 1. The comparison of the feature selection+stacking model and the original model

Model	Accuracy (%)		Time execution (s)	
	FS+Stacking model	Original model	FS+Stacking model	Original model
GBDT	92.5	91.95	70.68	14.28
AdaBoost	92.5	90.05	68.9	3.43
XGB	92.32	92.16	68.94	3.87
LGBM	92.46	92.45	66.41	0.7
DT	92.31	88.5	69.61	0.95
RF	92.54	92.16	70.11	8.74

4. CONCLUSION

This paper discusses the challenges of standard P2P lending data sets, such as high dimensions, small sample sizes, and unbalanced class labels. A feature selection technique based on the embedded technique is introduced. The most important features of the P2P lending data set were extracted within the framework with GBDT, RF, AdaBoost, XGBoost, LGBM, and DT. The result of feature selection is that each meta-learner in the embedded technique produces various important features based on the weight value of each feature so that the use of meta-learners greatly affects the accuracy score generated by the evaluation model. The stacking ensemble model produces an average of 94.54% accuracy and 69.10 s execution time. RF meta-learner+stacking ensemble is the best classification model, and LGBM meta- learner+stacking ensemble has the fastest execution time.

APPENDIX

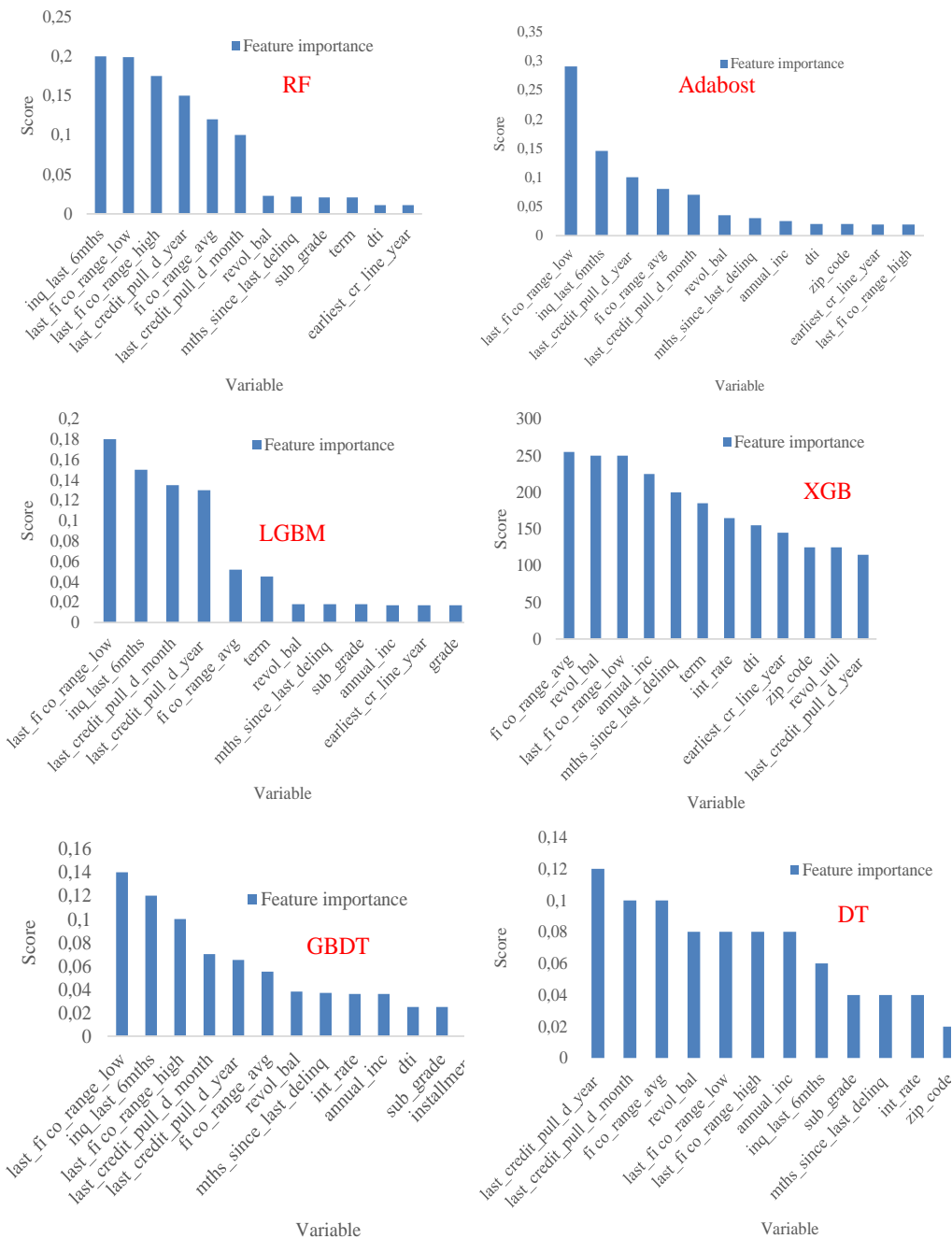


Figure 3. The features importance based on embedded technique





REFERENCES

- [1] WEF; Deloitte, *The Future of Financial Services-How disruptive innovations are reshaping the way financial services are structured, provisioned and consumed*, no. June. 2015.
- [2] S. F. Chen, G. Chakraborty, and L. H. Li, "Feature selection on credit risk prediction for peer to peer lending," in *JSAI International Symposium on Artificial Intelligence*, Springer, Cham, 2018, pp. 5–18, doi: 10.1007/978-3-030-31605-1_1.
- [3] L. C. Thomas, "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers," *International Journal of Forecasting*, vol. 16, no. 2, 2000, doi: 10.1016/S0169-2070(00)00034-0.
- [4] G. Attigeri, M. M. M. Pai, and R. M. Pai, "Framework to predict NPA/Willful defaults in corporate loans: A big data approach," *International Journal of Electrical & Computer Engineering*, vol. 9, no. 5, pp. 3786–3797, 2019, doi: 10.11591/ijece.v9i5.pp3786-3797.
- [5] H. Guo, K. Peng, X. Xu, S. Tao, and Z. Wu, "The prediction analysis of peer-to-peer lending platforms default risk based on comparative models," *Scientific Programming*, vol. 2020, pp. 1-10, 2020, doi: 10.1155/2020/8816419.
- [6] X. Yao, J. Crook, and G. Andreeva, "Support vector regression for loss given default modelling," *European Journal of Operational Research*, vol. 240, no. 2, pp. 528–538, 2015, doi: 10.1016/j.ejor.2014.06.043.
- [7] H. Kvamme, N. Sellereite, K. Aas, and S. Sjursen, "Predicting mortgage default using convolutional neural networks," *Expert Systems with Applications*, vol. 102, pp. 207–217, 2018, doi: 10.1016/j.eswa.2018.02.029.
- [8] M. J. Ariza-Garzón, J. Arroyo, A. Caparrini and M. -J. Segovia-Vargas, "Explainability of a machine learning granting scoring model in peer-to-peer lending," in *IEEE Access*, vol. 8, pp. 64873-64890, 2020, doi: 10.1109/ACCESS.2020.2984412.
- [9] Mukhtar *et al.*, "Hybrid model in machine learning–robust regression applied for sustainability agriculture and food security," *International Journal of Electrical & Computer Engineering*, vol. 12, no. 4, pp. 4457–4468, 2022, doi: 10.11591/ijece.v12i4.pp4457-4468.
- [10] S. Birla, K. Kohli and A. Dutta, "Machine learning on imbalanced data in credit risk," *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2016, pp. 1-6, doi: 10.1109/IEMCON.2016.7746326.
- [11] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446–3453, 2012, doi: 10.1016/j.eswa.2011.09.033.
- [12] M. Ashraf, G. Chetty, D. Tran, and D. Sharma, "Hybrid approach for diagnosing thyroid, hepatitis, and breast cancer based on correlation based feature selection and Naïve Bayes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Berlin, Heidelberg, 2012, vol. 7666 LNCS, no. PART 4, pp. 272–280, doi: 10.1007/978-3-642-34478-7_34.
- [13] Z. Yan and C. Yuan, "Ant colony optimization for feature selection in face recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2004, vol. 3072, pp. 221–226, doi: 10.1007/978-3-540-25948-0_31.
- [14] S. Wang and X. Yao, "Multiclass imbalance problems: analysis and potential solutions," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119–1130, Aug. 2012, doi: 10.1109/TSMCB.2012.2187280.
- [15] C. Wan and A. A. Freitas, "An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features," *Artificial intelligence review*, vol. 50, no. 2, pp. 201–240, 2018, doi: 10.1007/s10462-017-9541-y.
- [16] B. H. Shekar and G. Dagnev, "Classification of multi-class microarray cancer data using ensemble learning method," in *Lecture Notes in Networks and Systems*, Springer, Singapore, 2019, vol. 43, pp. 279–292, doi: 10.1007/978-981-13-2514-4_24.
- [17] V. Stanev *et al.*, "Machine learning modeling of superconducting critical temperature," *npj Computational Materials*, vol. 4, no. 1, 2018, doi: 10.1038/s41524-018-0085-8.
- [18] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179–189, 2018, doi: 10.1016/j.eij.2018.03.002.
- [19] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, pp. 1-26, 2020, doi: 10.1186/s40537-020-00327-4.
- [20] M. S. Sainin and R. Alfred, "A genetic based wrapper feature selection approach using nearest neighbour distance matrix," *2011 3rd Conference on Data Mining and Optimization (DMO)*, 2011, pp. 237-242, doi: 10.1109/DMO.2011.5976534.
- [21] J. K. Jaiswal and R. Samikannu, "Application of random forest algorithm on feature subset selection and classification and regression," *2017 World Congress on Computing and Communication Technologies (WCCCT)*, 2017, pp. 65-68, doi: 10.1109/WCCCT.2016.25.
- [22] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *Journal of Biomedical Science and Engineering*, vol. 06, no. 05, pp. 551–560, 2013, doi: 10.4236/jbise.2013.65070.
- [23] M. T. Uddin and M. A. Uddiny, "A guided random forest based feature selection approach for activity recognition," *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2015, pp. 1-6, doi: 10.1109/ICEEICT.2015.7307376.
- [24] S. Gharsalli, B. Emile, H. Laurent, and X. Desquesnes, "Feature Selection for Emotion Recognition based on Random Forest," *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP, (VISIGRAPP 2016)*, 2016, vol. 4, pp. 610–617, doi: 10.5220/0005725206100617.
- [25] Thanh-Tung Nguyen, J. Z. Huang, and T. T. Nguyen, "Unbiased feature selection in learning random forests for high-dimensional data," *The Scientific World Journal*, vol. 2015, pp 1-18, 2015, doi: 10.1155/2015/471371.
- [26] R. Wang, "AdaBoost for feature selection, classification and its relation with SVM, a review," *Physics Procedia*, vol. 25, pp. 800–807, 2012, doi: 10.1016/j.phpro.2012.03.160.
- [27] P. Poongothai and T. Devi, "Discriminant pearson correlative feature selection based gentle Adaboost Classification for medical document mining," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 3, pp. 3777–3783, 2019, doi: 10.35940/ijrte.c5391.098319.
- [28] M. F. Tolba and M. Moustafa, "GAdaBoost: accelerating adaboost feature selection with genetic algorithms," in *Proceedings of the 8th International Joint Conference on Computational Intelligence - ECTA, (IJCCI 2016)*, vol. 1, pp. 156–163, 2016, doi: 10.5220/0006041101560163.
- [29] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5549–5557, 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.





- [30] D. D. Li, D. X. Yu, D. X. Yu, Z. J. Qu, Z. J. Qu, and S. H. Yu, "Feature selection and model fusion approach for predicting urban macro travel time," *Mathematical Problems in Engineering*, vol. 2020, pp. 1-13, 2020, doi: 10.1155/2020/6897965.
- [31] E. Martin *et al.*, "Stacked generalization," *Encyclopedia of Machine Learning*, pp. 912-912, 2011, doi: 10.1007/978-0-387-30164-8_778.

BIOGRAPHIES OF AUTHORS







Muhammad Munsarif     received the Master Degree in Computer science from Dian Nuswantoro University (UDINUS) in 2002. Currently, he is a lecturer in informatics Engineering at Muhammadiyah University, Semarang (UNIMUS). His research interests include computer vision, data science and technopreneurship. He can be contacted at email: m.munsarif@unimus.ac.id.



Muhammad Sam'an     received Bachelor Degree from Universitas Negeri Semarang and Master Degree from Universitas Diponegoro in Mathematics 2010 and 2016 respectively. His research interests are in optimization, fuzzy mathematics and computational mathematics. He can be contacted at email: muhammad.92sam@gmail.com.



Safuan     received the Master Degree in Informatics Engineering from Dian Nuswantoro University (UDINUS) in 2015. Currently, he is a lecturer in informatics Engineering at Muhammadiyah University, Semarang (UNIMUS). His research interests include data mining, programming and web security. He can be contacted at email: safuan@unimus.ac.id.