A Blending Ensemble Approach to Predicting Student Dropout in Massive Open Online Courses (MOOCs)

Muhammad Ricky Perdana Putra[1], Ema Utami[2*]

[1,2]*Master of Informatics, Universitas Amikom Yogyakarta, Indonesia*

*corr-author: ema.u@amikom.ac.id

Abstract - The problem faced in the implementation of Massive Open Online Course (MOOC) is the high dropout rate (DO) reaching 90% which exceeds the formal school dropout rate. Preventive action needs to be taken to minimize the impact on MOOCs, instructors, and students. One solution is to do machine learning (ML) based prediction. The use of ML does not escape the problem of prediction performance that is still less accurate so it needs to be improved by blending ensemble learning (BEL). This research builds a BEL model consisting of two layers including base model with KNN, Decision Tree, and Naïve Bayes algorithms, then meta model with XGBoost. The dataset from KDD Cup 2015 contains clickstream from XuetangX website. The pre-processing stage includes selecting the course with the most participants, normalization, SMOTE, feature selection, and breaking it into three: ensemble, blender, and test data. The BEL model evaluation results obtained an accuracy value of 90.16%, precision of 85.64%, recall of 97.31%, F1-Score of 91.10%, and AUC of 92.83%.

Keywords: blending ensemble learning; MOOC; prediction; dropout; SMOTE.

I. INTRODUCTION

One of the problems faced by Massive Open Online Course (MOOC) organisers is the high dropout rate (DO) which reaches 90% [1]. The various causes of students dropping out of MOOCs are lack of social support, motivation, and perseverance [2], difficulty understanding the material, lack of engagement and interaction with the instructor [3], lack of understanding of learning goals and intentions [4], lack of support from peers [5], and the absence of adaptive variations in the learning flow so that students feel bored [6].

The impact of students dropping out will affect at least three parties: the organiser, the instructor, and the students themselves. MOOC organisers are impacted in terms of reputation, ranking, accreditation, as well as revenue [7]. On the instructor side, it affects motivation in teaching and developing new classes or materials as well as reduced social interaction [8]. Students who drop out may experience difficulties in finding employment and income in the future, worsening economic and social conditions [2].

One of the preventive efforts to deal with the problem of a large number of dropout students is to make early predictions with machine learning (ML). The challenge of this solution is the absence of direct supervision by instructors on students because MOOCs are held online [9]. What can be used as a basis for prediction is the student's interaction with the MOOC platform recorded in the activity log or clickstream which significantly influences the student's performance while attending the MOOC class [10].

Previous research developing ML models using single algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Naïve Bayes, Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) conducted by [1], [11], [12], has four problems namely lack of prediction performance, model generalisation, using only one dataset and one model testing technique. So it is necessary to optimise the model with optimisation techniques or ensemble learning by combining more than one algorithm.

One example of an optimisation technique is the Particle Swarm Optimisation (PSO) algorithm. Jin proposed a combined model between Support Vector Regression (SVR) and Improved Quantum PSO (IQPSO) called SVRQ [13]. The model is applied from week 2 to week 5. The accuracy value obtained is 87.16%, Area Under Curve (AUC) value 85.44%, and F1-Score 91.32%. The weaknesses of the study are the use of limited datasets and may cause a lack of generalisation of the model and only use one testing technique.

Another technique that can be applied is ensemble learning with types including boosting, bootstrap aggregating (bagging), stacking, and blending [14]. The ensemble learning technique can improve prediction performance and can overcome overfitting [15].

Ensemble stacking and blending have the same concept which consists of two layers, the first layer consists of a weak learner algorithm with the aim of converting complex non-linear features and the second layer is called a meta learner to utilise the residue of the first layer so as to improve the prediction model [16].

Research using ensemble learning was conducted by [17] with a combined model of Student Interaction Graph (SIG) and Neural Network or abbreviated as Sig-Net. The research compared the robustness of the Sig-Net model against two datasets sourced from KDD Cup and NAVER. The evaluation results are with the KDD Cup dataset getting an AUC value of 88.87% and F1-Score 92.52%, while NAVER gets an AUC value of 88.16% and F1-Score 97.28%. The weakness of this research is the lack of explanation of the pre-processing stage.

Another study conducted by [18] proposed the Ensemble Deep Learning Network (EDLN) model which is a combination of Faster RCNN and ResNet-50. The dataset came from KDD Cup 2015 and selected the first five weeks and then made multidimensional data. The data was mapped with metrics. Based on the evaluation results, the lowest accuracy was obtained from 1x7 metrics with a value of 87.7% and the highest accuracy was obtained from 7x7 metrics with a value of 97.4%. The weakness of the research is that it does not explain what features are used for prediction.

Research by [19] experimented with Decision Tree and bagging models such as AdaBoost, XGBoost, and Gradient Boosting (GB). After testing, of the four models, the GB algorithm gets a precision value of 0.75 and an AUC of 86.63%. Then, clustering is done with K-Means Clustering and using Ant Colony Optimisation (ACO) algorithm. With ACO, learning materials can be customised to reduce dropout rates. However, the implementation of ACO and the accuracy value are not explained in detail.

Based on the description above, there are two core problems underlying the research, namely practical problems and research problems. The practical problem is the high dropout rate that causes adverse impacts on MOOC organisers, instructors, and students that must be managed. The previous research problem is related to models that are still less than optimal and have the potential to be refined, especially vital parts such as the pre-processing stage and the selection of algorithms to build models that have a major effect on prediction accuracy.

Therefore, the purpose of this research is to build a blending ensemble learning (BEL) model to predict dropout students and determine the robustness of the model that has been built on several datasets. The limitations of this research are: (1) using a dataset from KDD Cup 2015 which contains activity logs from the largest MOOC site from China, XuetangX; (2) the dataset is grouped by course and one course is selected for deeper analysis; and (3) this research is conducted for academic purposes and not for real implementation in MOOCs.

II. METHOD

A. Dataset

The dataset comes from KDD Cup 2015 which contains activity logs from China's largest MOOC, XuetangX. The dataset has been uploaded on the Kaggle.com site by contributor Anas Nofal and has been pre-processed by changing the activity log data from rows to frequency columns. The dataset has been split into two, namely training data totalling 180,713 and test data totalling 44,929. The dataset consists of 30 features including 22 features related to activity logs, 1 feature related to classification labels, and the rest related to participant personal data presented in Table I.

TABLE I
DATASET FEATURE

No	Feature	No	Feature
1	Username	16	Action Create Comment
2	Course ID	17	Action Create Thread
3	Session ID	18	Action Delete Comment
4	Unique Session Count	19	Action Delete Thread
5	Action Per Session	20	Action Load Video
6	Timestamps	21	Action Pause Video
7	Time Difference	22	Action Play Video
8	Truth	23	Action Problem Check
9	Action Click About	24	Action Problem Correct
10	Action Click Courseware	25	Action Problem Incorrect
11	Action Click Forum	26	Action Problem Get
12	Action Click Info	27	Action Problem Save
13	Action Click Progress	28	Action Reset Problem
14	Action Close Courseware	29	Action Seek Video
15	Action Close Forum	30	Action Stop Video

B. Pre-processing

The pre-processing stage is the foundation for good model performance. This research conducts several pre-processing sub-stages including (i) ensuring the data type is numeric because the research approach used is quantitative. If there is data in a row that cannot be converted to numeric then the row will be deleted. (ii) Ensuring there is no missing data (null). If there is missing data in a row then the row will be deleted. (iii) Delete duplicate data.

Next, (iv) filtered the data based on the course and selected one course with the code '30640014' and the number of participants as many as 3,813. The course dataset will be analysed to find out the features with the highest contribution when the model makes predictions. (v) Perform data normalisation. The type of normalisation chosen is Min Max Scaler because the dataset has a high value difference. Min Max Scaler scales the data to zero to one so that the distance between data is not too large and eliminates negative values [20].

Then, (vi) because the class distribution is still unbalanced and this can have an impact on the performance of the model built, it is necessary to balance it with the Synthetic Minority Oversampling Technique (SMOTE). SMOTE can produce a synthetic class form that is considered better than duplicating the minority class so as to avoid overfitting [21] and overcome oversampling [22]. Visually, the class distribution results of each dataset before and after the SMOTE technique are presented in Fig. 1 and Fig. 2.

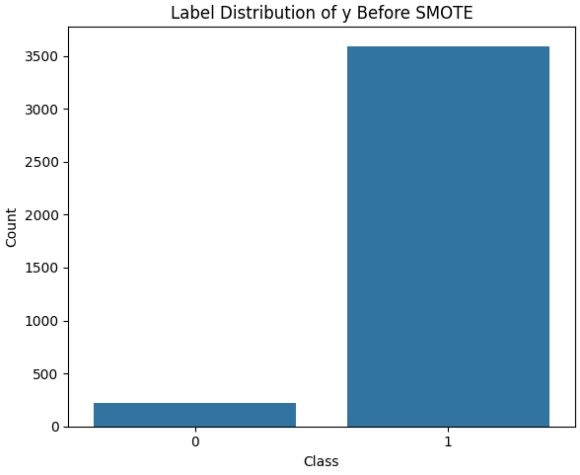

Fig. 1 Class distribution before SMOTE

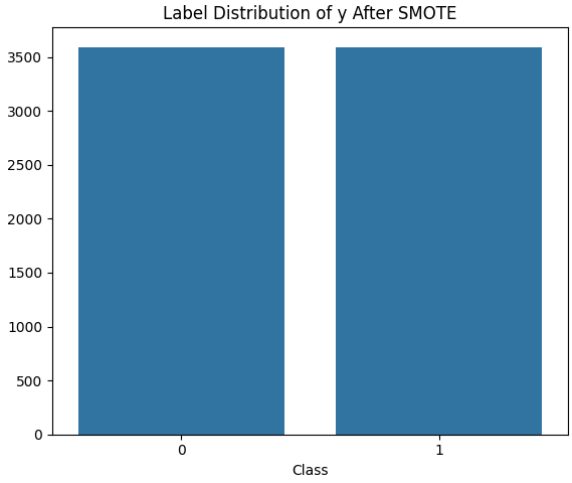

Fig. 2 Class distribution after SMOTE

Before the application of SMOTE, the dataset showed significant class imbalance with a ratio of 3593 DO classes and 220 no DO classes. After SMOTE was applied, the class distribution was successfully balanced, with both classes having the same number of 3593 rows. The total number of rows in the dataset increased to 7168. The application of SMOTE aims to overcome the problem of data imbalance which can improve the performance of the model in predicting both classes more accurately.

Finally, (vii) manually select features and use feature importance from Random Forest. Manually, only 22 features related to activity logs were selected and one classification label feature. Of the 22 features, a correlation analysis is carried out with a heatmap as in Fig. 3 so that it is known which features do not choose correlation with other features. This serves to compare the results of feature weighting with the Random Forest feature importance algorithm not contradicting each other.

From the heatmap, it can be seen that the 6th and 19th index features, namely action_close_forum and action_reset_problem, have no correlation with other features so that they are not used. The use of the feature importance algorithm can perform feature assessment and output in the form of contribution weights to predictions so as to improve model performance and reduce execution time [23]. The results of this stage are 15 features selected based on the highest weight to build the BEL model presented in Table II.

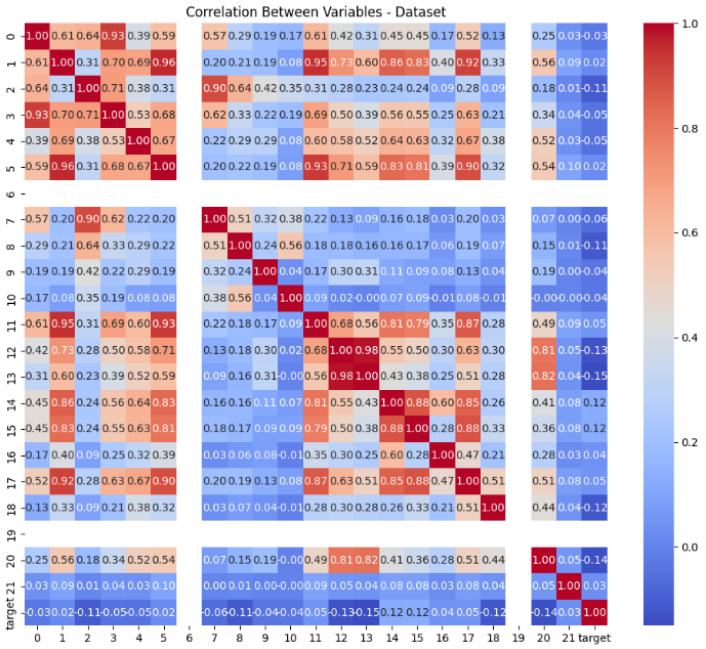

Fig. 3 Feature correlation

TABLE II
FEATURE SELECTION

Index	Feature	Index	Feature
21	Action Stop Video	15	Action Problem Check
17	Action Problem Get	18	Action Problem Save
16	Action Problem Incorrect	1	Action Click Courseware
0	Action Click About	13	Action Play Video
12	Action Pause Video	5	Action Close Courseware
11	Action Load Video	20	Action Seek Video
14	Action Problem Check	2	Action Click Forum
3	Action Click Info		

C. Model Building

Data that has been pre-processed is considered as mature data that is ready to be further processed with algorithms to produce prediction models with good performance. To build a BEL model requires three data, namely ensemble, blender, and test data. The split of the dataset with a percentage of 68%, 17%, and 15% as shown in Table III. The data used for model training is set to be larger to prevent overfitting, which is a condition where the model is too 'smart' with training data and not optimal on test data [22], [24].

The BEL model is built with two layers. The algorithms used in the first layer are called base models, namely KNN, Decision Tree, and Naïve Bayes. They were chosen because they are often used in previous research, have low accuracy values, are simple, and have the potential to be improved [25]. Training and testing on the first layer use ensemble and blender data. The result of the prediction or called the residue will be used as one data frame as training data for the second layer algorithm or called the meta model.

The algorithm in the meta model is XGBoost was chosen because it is able to utilise the residue from the previous model to build its own model so as to improve performance [16] and is able to handle class imbalance and scalability on large datasets [26]. After that, the BEL model will be tested using test data. The prediction results are used as the basis for calculating model performance. The flow of BEL model development refers to the research conducted by [27] and is visualized in Fig. 4.

TABLE III
DATASET SPLIT

Dataset	Split		
	Ensemble	Blender	Test
7.186	4.886	1.222	1.078

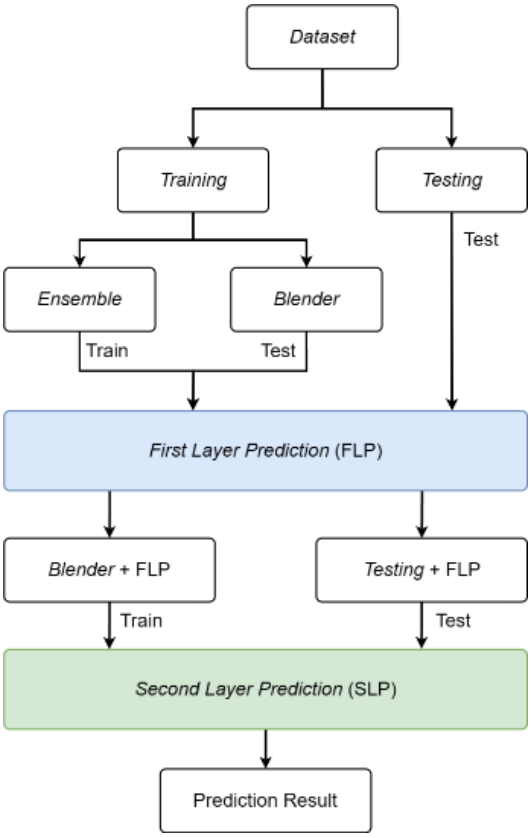

Fig. 4 Blending ensemble learning flow

D. Model Evaluation

Measurement of BEL model performance is done through three stages, namely (1) confusion matrix including accuracy, precision, recall, and F1-Score values. The results are displayed in visual form for easy analysis. (2) Using AUC to complete the confusion matrix. AUC is derived from Receiver Operating Characteristic (ROC) called Precision-Recall Curve [28] which has the advantage of not only focusing on finding the average accuracy but ROC visualises all possible classification thresholds [29], [30].

Then (3) uses k-fold cross validation where each iteration is calculated the confusion matrix value and AUC. The value of k chosen is 10 because it can provide a good balance between predictable error and variation in model evaluation and computational time is more efficient than high k values [31]. With k=10, the data will be divided into ten subsets, nine subsets as training data and one subset as test data. This process is carried out alternately until each subset has been used as test data.

III. RESULT AND DISCUSSION

This research involves a dataset by filtering it by course and selecting one of the courses. The research flow is as described previously starting from literature study, data collection, pre-processing, splitting the dataset and selecting algorithms for model building, and conducting test evaluation. After conducting experiments, the results are visually shown with the confusion matrix presented in Fig. 5 and the results of metric calculations are presented in Table 4.

The accuracy value shows 90.16%, which means that the model has a good ability to classify the data correctly, both for the DO and non-DO classes. Precision shows the level of accuracy of the model in predicting the positive class. The value of 85.64% means that the model can correctly predict the DO class. Then, the recall value of 97.31% indicates the model's ability to find all correct examples of the positive class. F1-Score is a combined metric of precision and recall, a value of 91.10% indicates a balance between the two.

The AUC value of 92.83% indicates a good ability of the model to distinguish between positive and negative classes, a good level of generalisation of the model, and the model is able to predict effectively on data that has never been seen before. The ROC curve presented in Fig. 6 provides a visual representation of the trade-off between false positive rate and true positive rate at various prediction thresholds. The larger the area under the ROC curve, the better the model is at handling the imbalance between positive and negative classes.

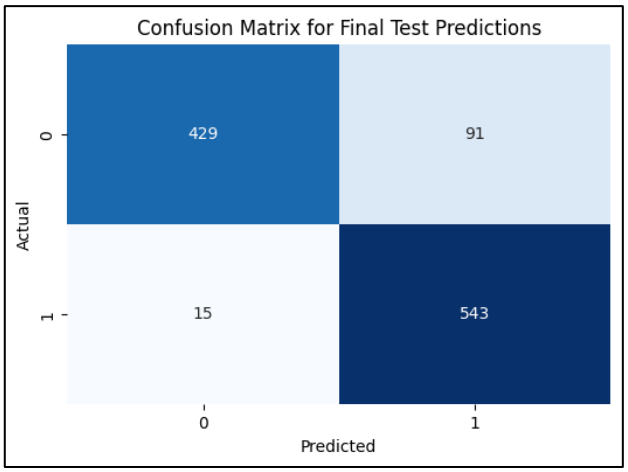

Fig. 5 Confusion matrix

TABLE IV
MODEL EVALUATION WITH CONFUSION MATRIX

Accuracy	Precision	Recall	F1-Score	AUC
90,16%	85,64%	97,31%	91,10%	92,83%

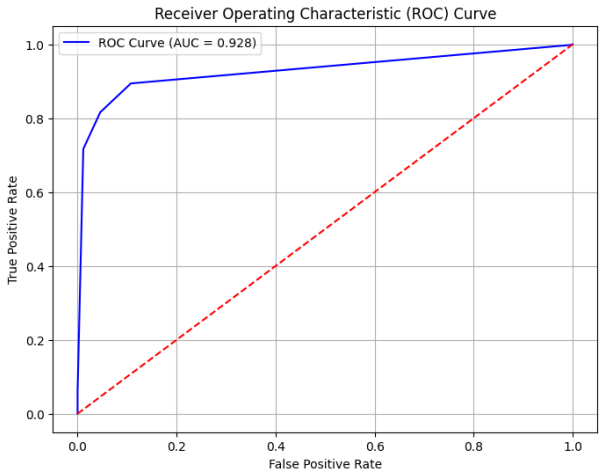

Fig. 6 Receiver Operating Characteristic (ROC)

Next, evaluate model testing using the k-fold cross validation technique. This aims to test the robustness of the model against 10 subsets of data. The test results are presented in tabulated form in Table V for easy analysis and reading. The data shows the consistency of model performance in k-fold testing although there are two iteration index that experience significant increases or fluctuations in value, namely the 5th and 6th index. Despite the fluctuations, the overall model still shows good stability.

These results show the consistency of the BEL model performance at k-fold. Although there were fluctuations in the 6th and 8th iterations. Both get an accuracy value of 90.18% and 96.63% where the average accuracy value is 89.60%. Another fluctuation occurred in the precision metric at the 5th iteration with a value of 94.62%. The recall metric did not escape the fluctuation at the 6th iteration with a value of 97.67%. In F1-Score, fluctuations occurred in the 6th and 8th iterations with values of 91.98% and 91.39%. AUC fluctuations occur at the same iteration as Accuracy and F1-Score with values of 95.23% and 95.00%.

The average fluctuation occurred at the 6th and 8th iterations on three metrics namely accuracy, F1-Score, and AUC. The analysis showed that both iterations had fewer zero values and the distance between data was not high enough. On the other hand, although F1-Score is a combination of precision and recall, it did not fluctuate at the same iteration. This shows that they do not necessarily affect each other. Then, the instability of AUC indicates the sensitivity of the model to the distribution of data in the training subset.

Research conducted [13] proposing the SVRQ ensemble model obtained an accuracy value of 87.16%, AUC 85.44%, and F1-Score 91.32%. The dataset used is limited to week 2 to week 5 only. While the research conducted does not limit the use of datasets within a certain time range, but focuses on one course. In terms of performance, the BEL model gets higher accuracy, AUC, and F1-Score values than SVRQ, namely 90.16%, 92.83%, and 91.10%. This proves that the BEL model is more optimal than SVRQ.

Comparison with previous research conducted by [17] shows that the research does not explain the pre-processing stage completely. Whereas, this research tries to cover the shortcomings by describing the pre-processing stage in as much detail as possible. In terms of performance, the BEL model is superior to the AUC metric. In addition, the testing metrics in this study are also more complete with confusion matrix, AUC, and k-fold cross validation. This gives a clearer picture of the effectiveness of the BEL model.

Another study by [18] with the EDLN model did get the highest accuracy value of 97.4%. However, the study used metric-based multidimensional data while this study uses single data so that the complexity of the dataset and model is certainly simpler. This research tries to complement the shortcomings of the research by describing in more detail what features are used for prediction along with multiple feature selection mechanisms, namely manually and feature selection algorithms.

TABLE V
K-FOLD CROSS VALIDATION

Fold	Accuracy	Precision	Recall	F1-Score	AUC
1	89,05%	84,16%	97,13%	90,18%	92,83%
2	89,42%	85,12%	96,41%	90,42%	92,35%
3	89,79%	85,78%	96,23%	90,70%	93,46%
4	89,05%	84,70%	96,23%	90,10%	94,21%
5	88,31%	94,62%	82,07%	87,90%	92,29%
6	90,18%	86,92%	97,67%	91,98%	95,23%
7	89,88%	85,57%	96,77%	90,83%	92,85%
8	90,63%	87,15%	96,05%	91,39%	95,00%
9	89,70%	85,64%	96,23%	90,63%	91,95%
10	88,96%	84,35%	96,59%	90,05%	92,59%

Research conducted by [19] comparing four Decision Tree-based models in the first layer for predicting dropout students. The second layer uses K-Means Clustering. The result of the prediction is used as the basis of ACO algorithm for learning path customisation. However, the ACO implementation and model performance are not explained in detail. The current research does not use clustering and ACO and focuses on improving prediction performance with the BEL concept.

From the four previous studies, some weaknesses are tried to be improved in this research, such as completing the pre-processing stage, feature selection, improving prediction performance, and completing testing techniques. On the other hand, this research still leaves room for further research such as the application of optimisation techniques such as PSO, ACO, and Komodo Mlipir Algorithm (KMA). Then, can experiment with clustering algorithms in the second layer, feature selection using genetic algorithms to make it more dynamic.

IV. CONCLUSION

The results prove that applying the BEL concept can improve the performance of the prediction model. This is linear with the four studies that are the main references and studies. Although there are some differences in them such as the source and treatment of datasets, algorithms used, optimization techniques, and model testing techniques. This research seeks to improve the shortcomings of the four studies by completing pre-processing, improving performance, and using model testing techniques. The improvised BEL model obtained an accuracy value of 90.16%, precision value of 85.64%, recall 97.31%, F1-Score 91.10%, and AUC 92.83%. Testing using k-fold cross validation produces a fairly stable performance value with an average accuracy value of 89.60%, precision 86.40, recall 95.14%, F1-Score 90.42%, AUC value 93.28%. Although in the 6th and 8th iterations there are three metrics that experience fluctuations, namely accuracy, F1-Score, and AUC. This shows that the BEL model is sufficiently generalized with the division of 10 subsets. Future research can explore other combinations of techniques or algorithms that can improve optimization such as ensemble learning, such as stacking, clustering algorithms, and dynamic feature selection. This is important to find a more efficient and effective approach related to data complexity. In addition, to know that the BEL model built has good generalization and robustness, it can compare the use of multiple datasets as done by [17].

REFERENCES

[1] Z. Chi, S. Zhang, and L. Shi, "Analysis and Prediction of MOOC Learners' Dropout Behavior," *Applied Sciences (Switzerland)*, vol. 13, no. 2, Jan. 2023, doi: 10.3390/app13021068.

[2] F. Agrusti, G. Bonavolontà, and M. Mezzini, "University dropout prediction through educational data mining techniques: A systematic review," *Journal of E-Learning and Knowledge Society*, vol. 15, no. 3, pp. 161–182, Oct. 2019, doi: 10.20368/1971-8829/1135017.

[3] W. Wunnasri, P. Musikawan, and C. So-In, "A Two-Phase Ensemble-Based Method for Predicting Learners' Grade in MOOCs," *Applied Sciences (Switzerland)*, vol. 13, no. 3, Feb. 2023, doi: 10.3390/app13031492.

[4] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes, "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model," *Decis Support Syst*, vol. 135, Aug. 2020, doi: 10.1016/j.dss.2020.113325.

[5] A. Alamri, M. Alshehri, A. Cristea, F.D. Pereira, E. Oliveira, L. Shi, and C. Stewart, "Predicting MOOCs dropout using only two easily obtainable features from the first week's activities," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2019, pp. 163–173. doi: 10.1007/978-3-030-22244-4_20.

[6] A. K. Darmawan and M. Makruf, "KLIK: Kajian Ilmiah Informatika dan Komputer Deteksi Gaya Belajar Siswa SMA pada Virtual Based Learning Environment(VBLE) dengan Decision Tree C4.5 dan Naive Bayes," *Media Online*, vol. 3, no. 5, pp. 532–544, 2023, [Online]. Available: https://djournals.com/klik

[7] L. J. Rodríguez-Muñiz, A. B. Bernardo, M. Esteban, and I. Díaz, "Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?," *PLoS One*, vol. 14, no. 6, Jun. 2019, doi: 10.1371/journal.pone.0218796.

[8] H. Huang, L. Jew, and D. Qi, "Take a MOOC and then drop: A systematic review of MOOC engagement pattern and dropout factor," Apr. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2023.e15220.

[9] H. Aldowah, H. Al-Samarraie, A. I. Alzahrani, and N. Alalwan, "Factors affecting student dropout in MOOCs: a cause and effect decision-making model," *J Comput High Educ*, vol. 32, no. 2, pp. 429–454, Aug. 2020, doi: 10.1007/s12528-019-09241-y.

[10] H. S. Park and J. Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning," 2021. [Online]. Available: www.joiv.org/index.php/joiv

[11] S. Nithya and S. Umarani, "MOOC Dropout Prediction using FIAR-ANN Model based on Learner Behavioral Features." [Online]. Available: www.ijacsa.thesai.org

[12] M. Şahin, "A Comparative Analysis of Dropout Prediction in Massive Open Online Courses," *Arab J Sci Eng*, vol. 46, no. 2, pp. 1845–1861, Feb. 2021, doi: 10.1007/s13369-020-05127-9.

[13] C. Jin, "MOOC student dropout prediction model based on learning behavior features and parameter optimization," *Interactive Learning Environments*, vol. 31, no. 2, pp. 714–732, 2023, doi: 10.1080/10494820.2020.1802300.

[14] T. Wu, W. Zhang, X. Jiao, W. Guo, and Y. Alhaj Hamoud, "Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration," *Comput Electron Agric*, vol. 184, May 2021, doi: 10.1016/j.compag.2021.106039.

[15] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, Jan. 2022, doi: 10.1016/j.caeai.2022.100066.

[16] N. I. Jha, I. Ghergulescu, and A. N. Moldovan, "OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques," in *CSEDU 2019 - Proceedings of the 11th International Conference on Computer Supported Education*, SciTePress, 2019, pp. 154–164. doi: 10.5220/0007767901540164.

[17] D. Roh, D. Han, D. Kim, K. Han, and M. Y. Yi, "SIG-Net: GNN based dropout prediction in MOOCs using Student Interaction Graph," in *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, Apr. 2024, pp. 29–37. doi: 10.1145/3605098.3636002.

[18] G. Kumar, A. Singh, and A. Sharma, "Ensemble Deep Learning Network Model for Dropout Prediction in MOOCs," 2023.

[19] E. M. Smaili, M. Daoudi, I. Oumaira, S. Azzouzi, and M. El Hassan Charaf, "Towards an Adaptive Learning Model using Optimal Learning Paths to Prevent MOOC Dropout," *International Journal of Engineering Pedagogy*, vol. 13, no. 7, pp. 128–144, 2023, doi: 10.3991/ijep.v13i7.40075.

[20] Henderi, T. Wahyuningsih , and E. Rahwanto, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," Mar. 2021. [Online]. Available: http://archive.ics.uci.edu/ml.

[21] M. Utari, "Implementation of Data Mining for Drop-Out Prediction using Random Forest Method," 2020.

[22] G. Psathas, T. K. Chatzidaki, and S. N. Demetriadis, "Predictive Modeling of Student Dropout in MOOCs and Self-Regulated Learning," *Computers*, vol. 12, no. 10, Oct. 2023, doi: 10.3390/computers12100194.

[23] F. Henni, B. Atmani, F. Atmani, and F. Saadi, "Improving Coronary Artery Disease Prediction: Use of Random Forest, Feature Importance and Case-Based Reasoning," *International Journal of Decision Support System Technology*, vol. 15, no. 1, 2023, doi: 10.4018/ijdsst.319307.

[24] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced mooc course using random forest model," *Information (Switzerland)*, vol. 12, no. 11, Nov. 2021, doi: 10.3390/info12110476.

[25] I. Y. Yunianto, M. M. M. Mutoffar, and A. K. Kurniawan, "Comparison of Decision Tree, KNN and Naïve Bayes Methods In Predicting Student Late Graduation In the Informatics Engineering Department, Institute Business XYZ," *Adpebi International Journal of Multidisciplinary Sciences*, vol. 1, no. 1, pp. 374–383, 2022, doi: 10.54099/aijms.v1i1.304.

[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[27] A. Hansrajh, T. T. Adeliyi, and J. Wing, "Detection of Online Fake News Using Blending Ensemble Learning," *Sci Program*, vol. 2021, 2021, doi: 10.1155/2021/3434458.

[28] K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3182.

[29] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, Jun. 2024, doi: 10.1016/j.patter.2024.100994.

[30] F. Movahedi, R. Padman, and J. F. Antaki, "Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores," *Journal of Thoracic and Cardiovascular Surgery*, vol. 165, no. 4, pp. 1433-1442.e2, Apr. 2023, doi: 10.1016/j.jtcvs.2021.07.041.

[31] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Comput Stat*, vol. 36, no. 3, pp. 2009–2031, Sep. 2021, doi: 10.1007/s00180-020-00999-9.