GLOBAL INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

https://global-us.mellbaou.com/

Open Access

Cite this article: Cynthia, E. P., Samuri, S. binti M., Li, W. S., Arta, Y., Syafitri, N., & Yanto, F. (2024). Hyperparameter Optimization of Random Forest Algorithm Technique in Acute Coronary Syndrome Cases. Global International Journal of Innovative Research, 2(3). https://doi.org/10.59613/global.v2i3.102

Received: February, 2024
Accepted: March, 2024

Keywords:
Index Terms—Acute Coronary Syndrom; Hyperparameter Optimization; Random Forest Algorithm

Author for correspondence:
Eka Pandu Cynthia
e-mail: eka.pandu.cynthia@uin-suska.ac.id

Published by:

GLOBAL SOCIETY PUBLISHING

Hyperparameter Optimization of Random Forest Algorithm Technique in Acute Coronary Syndrome Cases

[1]Eka Pandu Cynthia, [2]Suzani binti Mohamad Samuri, [3]Wang Shir Li, [4]Yudhi Arta, [5]Nesi Syafitri,[6]Febi Yanto

[1,4,5,6]Department of Software Engineering and Smart Technology, Faculty of Computing and Meta-Technology, Sultan Idris Education University, Tanjong Malim Perak, Malaysia

[2,3]Data Intelligence and Knowledge Management Special Interest Group, Faculty of Computing and Meta-Technology, Sultan Idris Education University, Tanjong Malim Perak, Malaysia

Research using random forest hyperparameter optimization in the case of acute coronary syndrome allows us to obtain a more optimal prediction model, but we can find a gap assumption where the learning carried out by the model still shows symptoms of over-fitting, characterized by a fairly large gap between the training and cross-training processes. validation in the model evaluation process. The research that will be carried out will provide a more optimal prediction model and will not produce symptoms of overfitting of the model using optimization techniques for the hyperparameters in the random forest algorithm. After carrying out various scenarios and testing accuracy, precision scores and various combinations of hyperparameters in the random forest algorithm, it was concluded that the model with the best optimization had a split ratio of 90:10 with an accuracy level of 84.44%, a precision score of 85, 29% and an MSE score of 0.1556 with the results of a combination of random forest optimization hyperparameters using gridCV. The optimization model using random grid cross validation that was built succeeded in reducing the level of over-fitting in the data, decreasing the MSE (mean squared error) from 0.17 and 0.24 to 0.15 for each model

© 2024 The Authors. Published by Global Society Publishing under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

1. Introduction

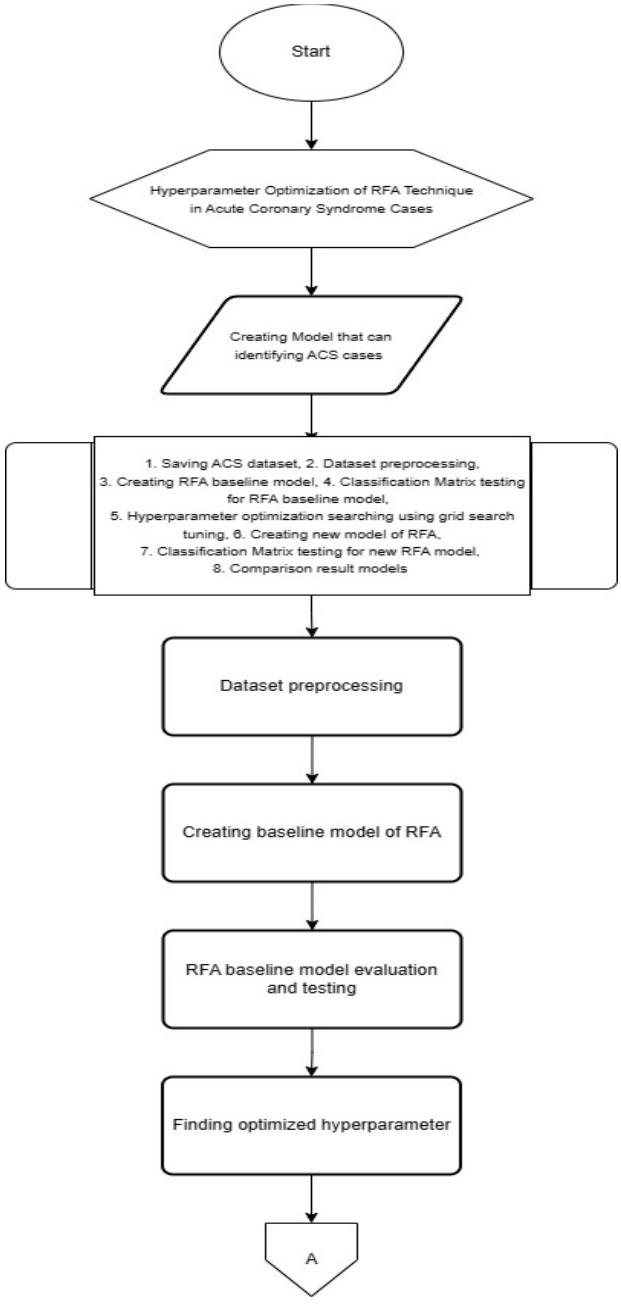
Heart disease is currently one of the highest causes of death in the world. Each year shows that 17.1 million people (29.1% of the total number of deaths) are caused by Coronary Syndrome Acute [1]. Indonesia is the country ranked first for heart cases with a percentage of 35% of the total heart cases in Southeast Asia [2]. Basic Health Research states that the national prevalence is 7.2% for heart disease and 8.7% for ischemic heart disease. Heart disease is also something that is receiving serious attention from the Indonesian Government, Minister of Health Nila F. Moeloek also touched on the issue of BPJS health funds which experienced a deficit for heart cases reaching Rp. 9.7 trillion (increased previously from IDR 9.5 trillion) which in this context consumes up to 30% of the health budget. The Social Security Administering Agency or BPJS Health noted that during 2018 it had spent IDR 79.2 trillion to pay claims for 84 million cases of illness of Indonesian citizens. The largest payment was given for claims for heart disease cases, namely IDR 9.3 trillion. BPJS Health Actuary Ocke Kurniandi explained that catastrophic illnesses or diseases requiring special treatment and high costs are the ones that burden the BPJS Health budget the most. For information at the household level, diseases identified as catastrophic diseases include cirrhosis hepatis, kidney failure, heart disease, cancer, stroke and blood diseases [2].

Research using random forest hyperparameter optimization in cases of acute coronary syndrome allows us to obtain a more optimal prediction model compared to not using optimization at all. The research that will be carried out is to continue a research entitled modeling using the random forest algorithm in cases of acute coronary syndrome [3] which resulted in the conclusion that using the default random forest algorithm in this case produced the best model with a data split ratio of 70% (70:30) stratified sampling with an accuracy level of 83.45%, precision 85% and recall 92.4%. However, in this research, a gap assumption was still found where the learning carried out by the model still showed symptoms of over-fitting, characterized by a fairly large gap between the training and cross-validation processes in the model evaluation process, in accordance with the conclusions given by the author. [3]. With the advice given by the author in previous research, namely by using optimization techniques on the hyperparameters in the random forest algorithm, it is hoped that the research that will be carried out will provide a more optimal prediction model and not produce symptoms of overfitting from the model. The resulting model is assessed using various statistical metric tables for classification cases so that the overall performance of the resulting model can be seen. This research is also expected to provide an understanding of the optimization techniques that will be carried out in optimizing the hyperparameters of the random forest

algorithm, and by using data science methodology, this research is also expected to provide a fairly clear picture regarding the patterns resulting from modeling acute coronary syndrome cases. so that both individuals and society can understand the cases being studied using various statistical images.

2. Research Method

The following is the methodology used in this research:



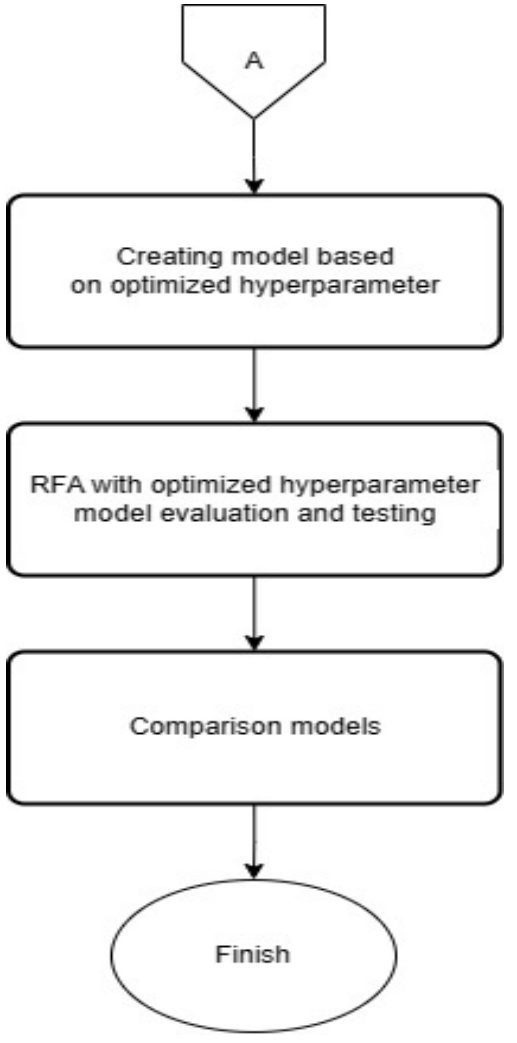


Fig. 1. Research Methodology

Based on the image above, what will be used will be explained one by one regarding the methodological design, namely:

A. Original Dataset

This stage is the stage of describing the dataset that has been used in previous research, namely acute coronary syndrome data taken from medical record sheets from IDI Medan City and the RSUD (Regional General Hospital) Arifin Achmad Pekanbaru. This dataset has 13 parameters that represent cases of acute coronary syndrome laboratory-wise and totals 444 pieces of data in the form of structural tables.

B. Preprocessing

In this process, we will search the data for statistical values and clean the data from various ambiguities, missing values, duplicate values, transform categorical data into dummies data

[8], in this process we will also split the data with a ratio of 70:30, 80:20, 90 :10 with a statistical separation technique, namely stratified sampling.

C. Random Forest Classification

At this stage, default modeling will be carried out first on the data which has been separated into training data and testing data using the random forest algorithm. The random forest process itself has a forest formation stage with a number of classification trees that can be determined where each decision tree formed is obtained from data formation using the random bootstrap method with random and random attribute candidates [9]. The bootstrap of the data is formed by dividing the sample by two-thirds of the length of the data and is formed according to the length of the data used, where data that is not included in the random bootstrap is used to calculate the amount of error in each tree that has been built [10]. The following is a flowchart and procedure for the random forest algorithm.

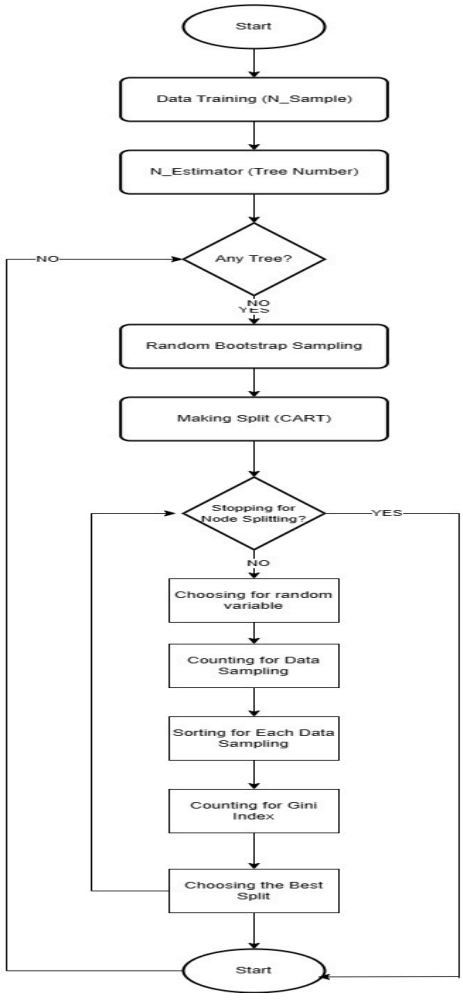


Fig. 2. Random Forest Flowchart

D. The Candidate of Hyperparameter Tuning Random Forest

The random forest algorithm has several hyperparameters that can be adjusted manually, where by adjusting the existing hyperparameters it will help to control model performance in terms of bias and variance, where the random forest algorithm itself should have low bias and high variance [11]. The candidate hyperparameters that can be used for optimization are as follows:

- Max_depth: is a hyperparameter that works to grow trees from the depth of the tree. If not controlled, the tree grows to the deepest depths.
- Min_sample_split: is a hyperparameter that determines the minimum number of samples required to divide internal nodes. Default value = 2.
- N_estimator: is a hyperparameter that determines the number of trees formed.
- Max_feature: is a hyperparameter that determines the maximum number of features used for the node splitting process. Type: sqrt, log2. If total features are n_features then: sqrt(n_features) or log2(n_features) can be selected as max features for node splitting.

E. Grid Search Tuning

Machine learning models have many hyperparameters to set and by tweaking these hyperparameters [12], the performance of the model can be improved. Hyperparameter tuning is the best method to perform various combinations of hyperparameters to assess classifier performance [13]. At this stage, after the default model of the random forest is formed, the optimization process using the grid search tuning technique will be implemented by combining various hyperparameters available by the random forest algorithm so that it can be seen which value of the parameter combination has better performance compared to the model without optimization [14]. The following is a flowchart of grid search tuning:

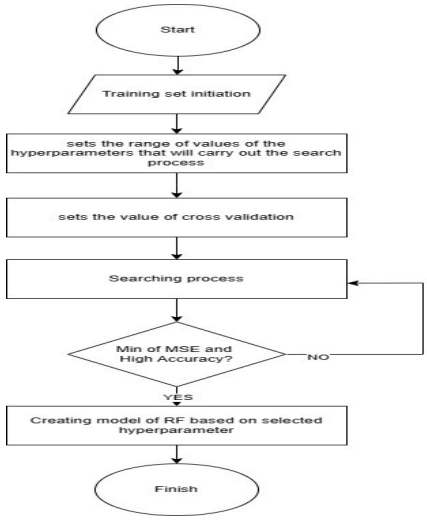


Fig. 3. Grid Search Tuning Flowchart

F. Model Evaluation

At this stage, testing will be carried out on the model that has been built by providing a comparison between the default random forest model from previous research in order to determine the level of success and performance of the random forest algorithm in building a model with optimization using the CV grid search technique for cases of acute coronary syndrome [6]. The comparison table will use statistical metrics to assess a classification case with the following details:

- Accuracy testing, is a testing method used to calculate the level of accuracy with a confusion matrix table which is depicted in a table that states the number of correct test data and incorrectly classified test data [15].
- Precision and recall testing is a testing technique used to calculate performance using the confusion matrix table of the algorithm used in a case [16]. Precision is used to calculate the true positive ratio compared to the overall predicted positive results. Meanwhile, recall is the ratio of true positive predictions compared to the total true positive data (True positive ratio) [17].

3. Result and Discussion

The implementation in this research includes: modeling built using the Python programming language, this modeling does not use a database, but rather accesses it data file in CSV form on the computer's operating system. Dataset models for comparison are 80:20 (80%), 70:30 (70%), 90:10 (90%). The stages that will be carried out will be explained as follows.

A. Random Forest Classification Baseline Model

At this stage, initiate a random forest model as a baseline model before we carry out optimization techniques with grid search tuning. We will determine the parameters of the number of trees (n_estimator) that will be built and the depth level. In this research, a baseline model was built with a depth of 4 and 100 trees were built and the model will train on training data at a ratio of 70:30, 80:20, and 90:10.

B. Sample Tree in Forest Visualization

The next stage will display one tree sample from a ratio of 80:20 from the many trees in the forest.

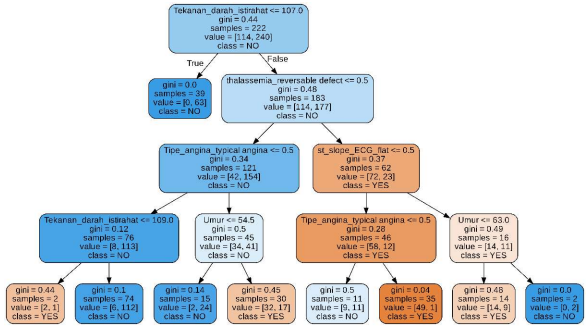


Fig. 1. Sample Tree in Baseline Forest

C. Random Forest Baseline Evaluation

At this stage, we will explain the evaluation of the baseline model produced by the default random forest with metric statistical evaluation for classification (Accuracy, Precision, and Recall) at each split ratio that has been carried out so that the overall performance of the random forest baseline can be compared with the metrics. evaluation when carrying out gridsearch tuning at the next stage. The following is an evaluation table produced from the baseline random forest model:

Table 1 Random Forest Baseline Clasifier Evaluation

Ratio Split Baseline Model	**Evaluation**			
	Accuracy	***Precision***	***Recall***	***MSE***
Ratio 70:30	82, 70 %	84, 1 %	92, 4 %	0.1729
Ratio 80:20	82, 02 %	80, 8 %	96, 7 %	0.1797
Ratio 90:10	75, 55 %	83, 3 %	96, 8 %	0.2444

D. K-Fold Cross Validation Baseline Evaluation

At this stage, we will explain the baseline model evaluation using basic k-fold cross validation testing of 7 folds to see overall how the model works from the data used for each split ratio as follows:

Table 2 Random Forest Baseline Clasifier Evaluation

Ratio Split Baseline Model	***Fold***	***Accuracy***
Ratio 70:30	7 fold	84, 9 %
	10 fold	87, 2%
	11 fold	85, 4%
Ratio 80:20	7 fold	84, 9 %
	10 fold	85, 6%
	11 fold	85, 6%
Ratio 90:10	7 fold	84, 9 %
	10 fold	85, 6 %

	11 fold	85, 6%

After we know the baseline cross-validation we will see how the baseline model works on the data used by looking at the validation curve and learning curve for each as follows:

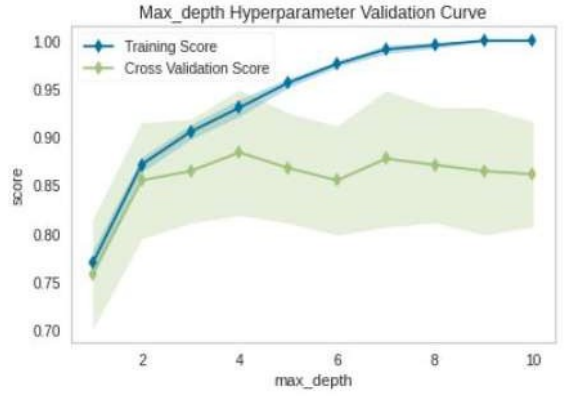


Fig. 2. Validation Curve for Ratio 70:30

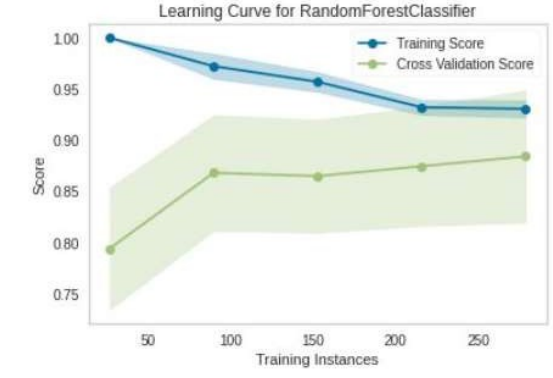


Fig. 3. Learning Curve for Ratio 70:30

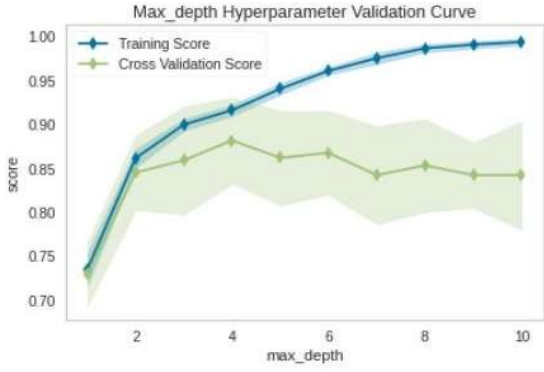


Fig. 4. Validation Curve for Ratio 80:20

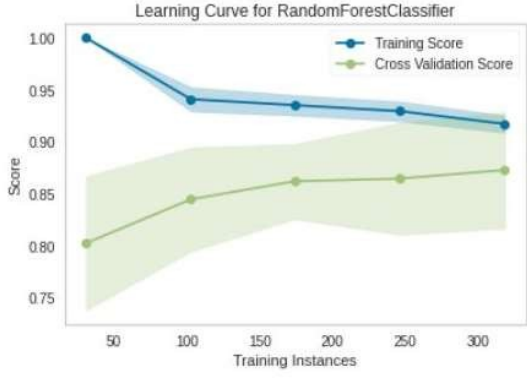


Fig. 5. Learning Curve for Ratio 80:20

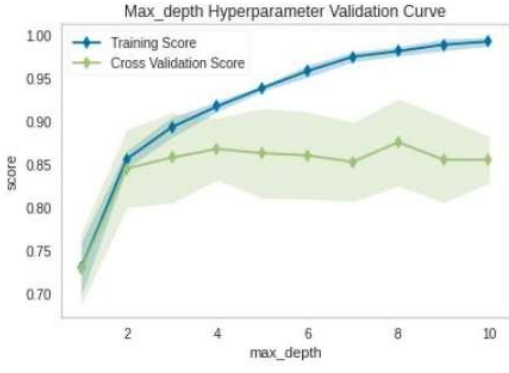


Fig. 6. Validation Curve for Ratio 90:10

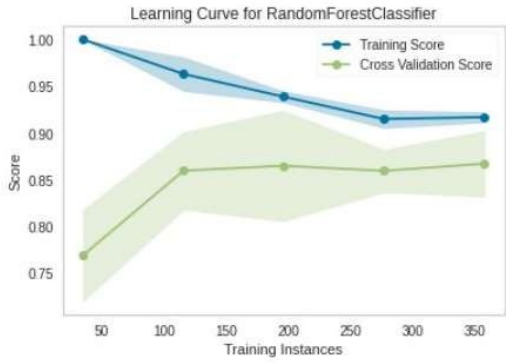


Fig. 7. Learning Curve for Ratio 90:10

From the three curves above, in the baseline model, each ratio still shows symptoms of over-fitting to data at a deeper tree level as shown by the validation curve at ratios 70:30 and 80:20 which tends to decrease and is marked by the MSE (mean) value. squared error) which still tends to be high in the range of 0.17 at a ratio of 70:30 and 80:20 and 0.24 at a ratio of 90:10, therefore it is necessary to carry out optimization to reduce the level of this tendency.

Grid Search Hyperparameter Tuning

At this stage, we will explain the evaluation of the baseline model produced by the default random forest. In the first stage of Grid Search Tuning Hyperparameters, we will initiate range values and parameters. We also need to declare the candidates for the tuning that we will carry out so that we can later compare which candidates produce higher performance and what parameters and hyperparameters were selected from these candidates.

Each of the parameters and hyperparameters at input ratios of 70%, 80%, 90% above is trained using 7-fold training data and cross-validation and each parameter is looped 12 times to get maximum results so as to find performance based on the best accuracy and hyperparameters.

So it can be concluded that the random forest hyperparameter tuning technique has succeeded in improving performance compared to using the usual random forest algorithm, which can be explained in the following table:

Table 3 Random Forest vs Grid SearchCV Random Forest Accuracy Result

Baseline Random Forest	**Rasio 70**	**Rasio 80**	**Rasio 90**
	82,7 %	82, 02 %	75, 56%

Proposed Method (Random Forest + Grid SearchHyper ParameterTuning)	**84, 2%**	**84, 26%**	**84, 44%**

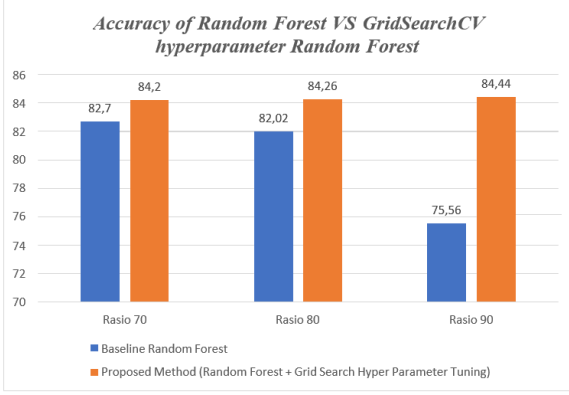


Fig. 8. Random Forest vs Grid SearchCV Random Forest Accuracy Result Diagram

Table 4 Random Forest vs Grid SearchCV Random Forest Precission Result

Baseline Random Forest	**Rasio 70**	**Rasio 80**	**Rasio 90**
	84,15 %	80, 82 %	77, 78%
Proposed Method (Random Forest + Grid Search Hyper ParameterTuning)	**85, 85%**	**85, 07%**	**85, 29%**

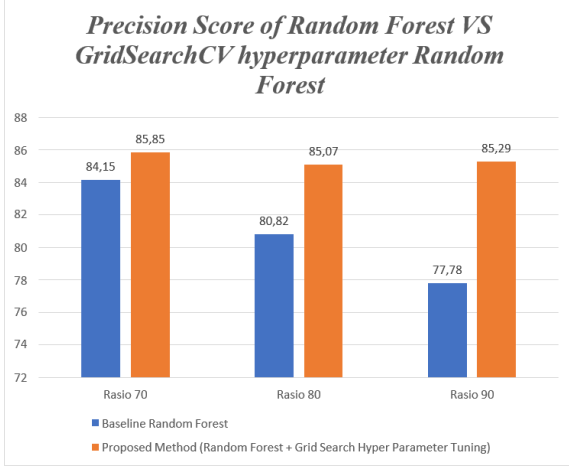


Fig. 9. Random Forest vs Grid SearchCV Random Forest Precission Result Diagram

Table 5 Random Forest vs Grid SearchCV Random Forest MSE Result

Baseline Random Forest	Rasio 70	Rasio 80	Rasio 90
	0,1729	0,1797	0,2444
Proposed Method (Random Forest + Grid Search Hyper Parameter Tuning)	**0,1578**	**0,1573**	**0,1556**

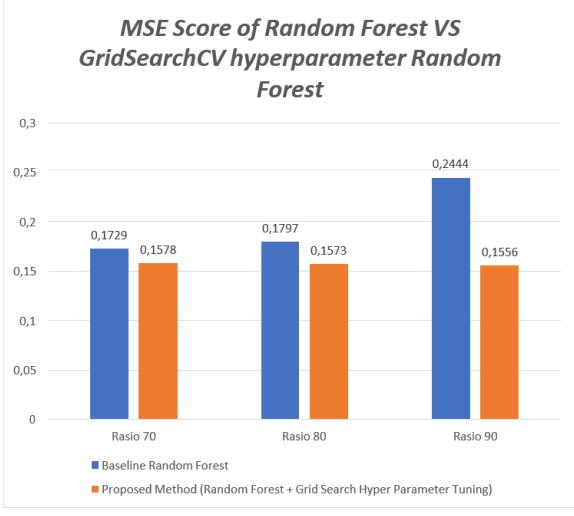


Fig. 10. Random Forest vs Grid SearchCV Random Forest MSE Result Diagram

4. Conclusion

Based on the stages that have been carried out in this research, conclusions can be drawn, namely:

- Hyperparameter optimization for random forest algorithm modeling using the random grid cross validation technique in the case of acute coronary syndrome has been successfully developed in accordance with the analysis and design that has been carried out.

- After carrying out various scenarios and testing accuracy, precision scores and various combinations of hyperparameters in the random forest algorithm, it was concluded that the model with the best optimization had a split ratio of 90:10 with an accuracy level of 84.44%, a precision score of 85 .29% and an MSE score of 0.1556 with the results of a combination of random forest optimization hyperparameters using gridCV being {'max_features': 'log2', 'n_estimators': 200}.

- In this study, the optimization model using random grid cross validation succeeded in reducing the level of over-fitting in the data, decreasing the MSE (mean squared error) from 0.17 and 0.24 to 0.15 for each model

5. References

[1] W. H. Organization, "Cardiovascular diseases," 2017. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Mar. 27, 2021).

[2] P. Fauzi, "Kata Menkes Soal Dana BPJS Kesehatan Rp 9 T untuk Penyakit Jantung," 2018. https://health.detik.com/berita-detikhealth/d-4288339/kata-menkes-soal-dana-bpjs-kesehatan-rp-9-t-untuk-penyakit-jantung (accessed Mar. 06, 2021).

[3] E. P. Cynthia, M. Afif Rizky A., A. Nazir, and F. Syafria, "Random Forest Algorithm to Investigate the Case of Acute Coronary Syndrome," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 2, pp. 369–378, 2021, doi: 10.29207/resti.v5i2.3000.

[4] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," IEEE Trans. Image Process., vol. 27, no. 8, pp. 4012–4024, Aug. 2018, doi: 10.1109/TIP.2018.2834830.

[5] Z. Masetic and A. Subasi, "Congestive Heart Failure Detection Using Random Forest Classifier," Comput. Methods Programs Biomed., vol. 130, pp. 54–64, 2016, doi: 10.1016/j.cmpb.2016.03.020.

[6] R. P. Kaur, M. Kumar, and M. K. Jindal, "Performance evaluation of different features and classifiers for Gurumukhi newspaper text recognition," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 8, pp. 10245–10261, 2023, doi: 10.1007/s12652-021-03687-8.

[7] C. G. Siji George and B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 9, pp. 173–178, 2020, doi: 10.14569/IJACSA.2020.0110920.

[8] A. M. Antoniadi, M. Galvin, M. Heverin, O. Hardiman, and C. Mooney, "Prediction of caregiver burden in amyotrophic lateral sclerosis: a machine learning approach using random forests applied to a cohort study," BMJ Open, vol. 10, no. 2, p. e033109, Feb. 2020, doi: 10.1136/bmjopen-2019-033109.

[9] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," WIREs Data Min. Knowl. Discov., vol. 9, no. 3, May 2019, doi: 10.1002/widm.1301.

[10] T. I. Rohan, Awan-Ur-Rahman, A. B. Siddik, M. Islam, and M. S. U. Yusuf, "A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost," in 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Jul. 2019, no. June

2020, pp. 1–4, doi: 10.1109/IC4ME247184.2019.9036697.

[11] X. Qian, “A review of methods for sleep arousal detection using polysomnographic signals,” Brain Sciences, vol. 11, no. 10. 2021, doi: 10.3390/brainsci11101274.

[12] C. Krittanawong, “Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection,” Sci. Rep., vol. 11, no. 1, p. 8992, 2021, doi: 10.1038/s41598-021-88172-0.

[13] 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), Feb. 2019, pp. 1–8, doi: 10.1109/ICACCP.2019.8882943.

[14] M. Daviran, A. Maghsoudi, R. Ghezelbash, and B. Pradhan, “A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach,” Comput. Geosci., vol. 148, p. 104688, Mar. 2021, doi: 10.1016/j.cageo.2021.104688.

[15] L. M. Fleuren et al., “Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy,” Intensive Care Med., vol. 46, no. 3, pp. 383–400, Mar. 2020, doi: 10.1007/s00134-019-05872-y.

[16] S. Ahmad, “Diagnosis of cardiovascular disease using deep learning technique,” Soft Comput., vol. 27, no. 13, pp. 8971–8990, 2023, doi: 10.1007/s00500-022-07788-0.

[17] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, “The area under the precision-recall curve as a performance metric for rare binary events,” Methods Ecol. Evol., vol. 10, no. 4, pp. 565–577, Apr. 2019, doi: 10.1111/2041-210X.13140.