IMBGAFS: GA FEATURE SELECTION FOR AUC IN BIRD STRIKE PREDICTION

Aji Gautama Putrada and Sidik Prabowo

[1]Advanced and Creative Networks Research Center, Telkom University, Bandung,
Indonesia
[2]School of Computing, Telkom University, Bandung, Indonesia

ABSTRACT

Several studies discuss airplane failure prediction due to bird strikes. However, these studies need to analyze further the imbalance in their dataset. Our research aim is to create an airplane failure prediction by bird strike using a machine learning method optimized using GA feature selection. GA feature selection uses AUC maximization as the objective function to tackle imbalance problems in the bird strike dataset. First, we obtained the airplane bird strike dataset from Kaggle. We carry out preprocessing on the dataset.We then compared and chose one of four state-of-the-art machine learning methods: SVM, MLP, logistic regression, and random forest. The selection process involves oversampling methods, synthetic minority oversampling technique (SMOTE), and optimum threshold selection, which involves geometric mean (g- mean) and area under curve (AUC) values. Finally, we optimize airplane failure prediction by performing AUC maximization using GA feature selection. Our test results show that random forest is the best machinelearning method in airplane failure prediction compared to SVM, logistic regression, and MLP. SMOTE can increase random forest AUC from 0.845 to 0.878. Finally, the random forest model from ImbGAFS is better than the conventional method without feature selection. The increase in the AUC value is from 0.878 to 0.889. Then, after carrying out optimal threshold selection, ImbGAFS+random forest also has better sensitivity, specificity, and g-mean than conventional methods. The increase is from 0.7737, 0.8350, and 0.8037 to 0.8033, 0.8301, and 0.8166, respectively.

KEYWORDS

Genetic Algorithm, Area Under Curve Maximization, Airplane Failure, Imbalanced Dataset, Bird Strike

1. INTRODUCTION

Birds are warm-blooded vertebrate living things, identifiable by their beaks, feathers, and flight abilities. They have a wide range of varieties found in ecosystems around the world [1]. These animals are widely known for their uses, including pest control and plant seed dispersal. They also have cultural and aesthetic values due to their shapes and songs [2]. However, on the other hand, birds also often bring bad impacts, namely as agricultural pests, spreaders of disease, and disrupt flights [3]. Airplane bird strike is a term to describe an airplane colliding with a bird or a group of birds, which can cause disruption to the aircraft and threaten the bird's safety.

There have been many studies that have also tried to predict airplane failure using machine learning. Celikmih et al. [4] predicted the number of airplane equipment failures using data such as the number of equipment dismantled, the duration the equipment has been used to fly, and the

David C. Wyld et al. (Eds): MLNLP, NWCOM, DTMN, ASOFT, SIGPRO, AIFZ, CSITY, CLSB - 2023
pp. 203-216, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131614

number of equipment that has been unplanned. That study used three prediction regression methods: multi-layer perceptron (MLP), support vector machine (SVM) regression, and linear regression. Lu *et al.* [5] applied the remaining useful life (RUL) of the aircraft engine using logistic regression. That method is supported by a multi-sensor prognostic model using the novel Kalman filter online sequential extreme learning machine (KFOS-ELM). In addition, Yan *et al.* [6] used random forest for aircraft engine fault diagnosis. Comparing MLP, SVM, linear regression, and random forest for airplane failure prediction caused by bird strikes is a research opportunity.

The optimum machine learning model selection performance can still be improved by feature selection, where the genetic algorithm (GA) is one of the superior feature selection methods. Toma *et al.* [7] used GA for feature selection for fault detection on induction machine bearings. The model used is a decision tree. Chui *et al.* [8] used GA to balance recurrent neural network (RNN) and long short-term memory (LSTM) in RUL on turbofan engines. Using GA feature selection for airplane failure due to bird strikes is a research opportunity.

Our research aim is to create an airplane failure prediction using a machine learning method optimized using GA feature selection. GA feature selection uses AUC maximization as the objective function. First, we obtained the airplane bird strike dataset from Kaggle. We carry out preprocessing on the dataset. We then compared and chose one of four state-of-the-art machine learning methods: SVM, MLP, logistic regression, and random forest. The selection process involves oversampling methods, synthetic minority oversampling technique (SMOTE), and optimum threshold selection, which involves geometric mean (g-mean) and area under curve (AUC) values. Finally, we optimize airplane failure prediction by performing AUC maximization using GA feature selection.

To the best of our knowledge, there is no airplane failure detection due to bird strikes that utilize GA and AUC maximization. The following is a list of our research contributions:

1. An airplane failure detection due to bird strike, which has optimum performance on an imbalanced dataset.
2. A method for dealing with imbalanced datasets that combines ROC threshold selection and AUC maximization.
3. ImbGAFS, a novel GA feature selection that uses ROC threshold selection as the objective function.

The remainder of this paper is structured systematically: Section 2 discusses comprehensively state-of-the-art papers in airplane failure prediction by bird strike. Then Section 3 shows the methodology and theory at each stage. Next, Section 4 shows and compares our test results with state-of-the-art papers. Lastly, Section 5 emphasizes the important findings in this paper.

2. RELATED WORKS

Several studies discuss airplane failure prediction due to bird strikes. Nimmagadda *et al.* [9] performed airplane failure prediction by comparing k-nearest neighbors (KNN), decision trees, and Gaussian naïve Bayes classification. Gaussian naïve Bayes had the best performance with an accuracy of 0.86. Misra *et al.* [10] also carried out airplane failure prediction, but with models, namely random forest, artificial neural network (ANN), logistic regression, SVM, and extreme gradient boosting (XGBoost). Random forest was again the model with the best accuracy, namely 0.79. However, these studies need to analyze further the imbalance in their dataset.

One of the optimization methods in the imbalanced dataset is the ROC threshold selection. Ojo *et al.* [11] used ROC threshold selection in fault detection on lithium-ion batteries. The machine learning method used in this research is ANN. Sadeghi *et al.* [12] leveraged the ROC threshold selection for imbalance problems in detecting diabetes mellitus. The classification model used is a deep neural network (DNN), while the oversampling method is repeated edited nearest neighbor (RENN). Using ROC threshold selection for imbalance problems in airplane failure prediction due to bird strikes is a research opportunity.

AUC maximization, in addition to ROC threshold selection, improves prediction performance on imbalanced datasets. Yan *et al.* [13] used AUC maximization using a stochastic gradient method called the stochastic primal-dual algorithm. Wang *et al.* [14] performed AUC maximization with proximal SVM (PSVM) on the imbalanced dataset in composite outcomes of hospitalized COVID-19 patients. The AUC results were better than 25 other maximization methods. Using AUC maximization with ROC threshold selection is a research opportunity.

Finally, the GA method is one of the most superior methods in feature selection. The use of GA in feature selection in Alawad's research *et al.* [15] improved the performance of the extra tree classifier, random forest, support vector machine, and KNN in identifying brain hemorrhage. In research by Yang *et al.* [16], GA is used in conjunction with the time-series feature extractor (Tsfresh) to extract the signal and select the best features from the IoT data stream for anomaly detection. The classification uses XGBoost. Using GA feature selection with g-mean threshold selection as the objective function is a research opportunity.

Table 1. Related Works Comparison

Reference	Airplane Failure byBird Strike	ROC Threshold Selection	AUC Maximization	GA Feature Selection
Nimmagadda et al. [9]	Yes	No	No	No
Misra et al. [10]	Yes	No	No	No
Ojo et al. [11]	No	Yes	No	No
Sadeghi et al. [12]	No	Yes	No	No
Yan et al. [13]	No	No	Yes	No
Wang et al. [14]	No	No	Yes	No
Alawad et al. [15][13]	No	No	No	Yes
Yang et al. [16][14]	No	No	No	Yes
Proposed Method	Yes	Yes	Yes	Yes

Table 1 is a table that compares state-of-the-art papers in airplane failure detection by bird strike with GA and AUC maximization. The table also compares these studies with our research to highlight our research contribution.

3. PROPOSED METHOD

We propose a research methodology to achieve our research aim. First, we obtain the airplane bird strike dataset from Kaggle. We carry out preprocessing on the dataset. We then compared and chose one of four state-of-the-art machine learning methods: SVM, MLP, logistic regression, and random forest. The selection process involves oversampling, SMOTE, and optimum threshold selection methods concerning g-mean and AUC values. Finally, we optimize airplane

failure prediction by performing AUC maximization using GA feature selection. Fig. 1 shows our research methodology as a flow chart.

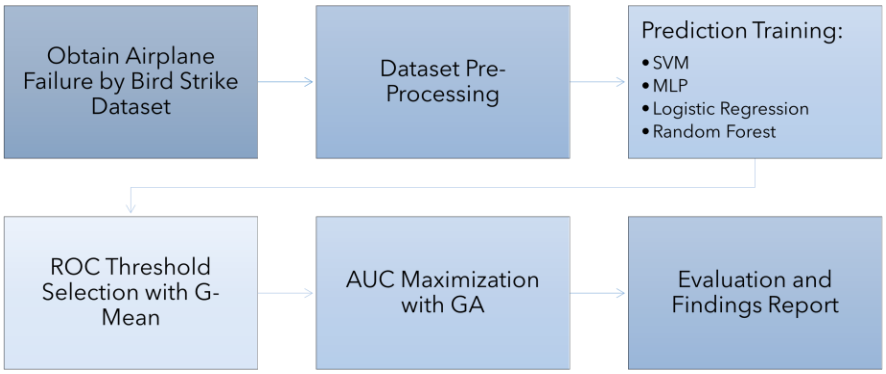


Figure 1. Our proposed methodology

3.1. Airplane Failure Detection by Bird Strike

Airplane bird strike is a term to describe an airplane colliding with a bird or a group of birds, which can cause disruption to the aircraft and threaten the bird's safety. "Miracle on The Hudson" is one of the famous bird strike events lately was made into a film called "Sully" (2016) with Tom Hanks as the titular role [17]. In the film, a plane taking off is hit by a bird strike and is forced to make an emergency landing in the Hudson River.

We obtain the bird strike dataset from Kaggle. The dataset is taken from voluntary bird strike data from the Federal Aviation Administration (FAA) [18]. The dataset consists of 23 columns and 65,611 data items. Of the 23 columns, we took the 19 most prominent columns. The explanation of these columns is as follows:

1. 'Airport: Name': The airport the flight departs from. There are 612 airport names in the dataset
2. 'Altitude bin': Range of aircraft altitudes. There are three ranges.
3. 'Aircraft: Make/Model': Aircraft model code. There are 255 aircraft model codes.
4. 'Effect: Impact to flight': The result of a bird strike. There are five consequences of a bird strike.
5. 'FlightDate': Flight date. The unordered start date is 1 January 2005, while the last is 12 December 2011.
6. 'Record ID': Record ID of the bird strike event. The smallest record ID is 15205, while the largest is 322935.
7. 'Effect: Indicated Damage': There are two values, "No damage" and "Caused damage."
8. 'Aircraft: Number of engines?': Number of engines on board. The range is 1 to 4.
9. 'Aircraft: Airline/Operator': Name of the aircraft carrier. There are 151 airline names.
10. 'Origin State': Origin state of the aircraft. There are 58 origin states in the dataset.
11. 'When: Phase of flight': In what phase does the bird strike occur? There are nine phases in the dataset
12. 'Wildlife: Size': Bird size. There are three sizes.
13. 'Wildlife: Species': The species name of the bird. There are 302 bird species names in the dataset.
14. 'When: Time (HHMM)': Hours and minutes the incident occurred. The range is from 00:00 to 23:59.

15. 'When: Time of day': There are four times of the day: "Dawn," "Day," "Dusk," or "Night."
16. 'Pilot warned of birds or wildlife?': The answer is yes or no.
17. 'Miles from airport': Distance from the nearest airport. Its range is 0 to 33.
18. 'Feet above ground': The range is 0 to 13,000.
19. 'Speed (IAS) in knots': The range is 0 to 300.

We use 'Effect: Indicated Damage' as a label and the others as features.

In the pre-processing stage, we use zero imputation to fill in the missing values in the dataset. Zero imputation fills the missing values with 0 [19]. Then, we use label encoding to convert categorical values into numerical values so that machine learning can process [20]. We perform feature analysis using the Pearson correlation to observe linear correlations between features [21]. Finally, we standardize the features so each feature has a similar range [22].

We benchmark this study's four machine learning models: random forest, SVM, MLP, and logistic regression. Random forest is part of the ensemble learning method using bootstrap and aggregating (bagging) [23]. The essence of the random forest is to conduct majority voting on several decision trees called weak learners for generalization purposes. SVM is a classification method that separates training data in feature space with a hyperplane [24]. The hyperplane is a field that can divide the data linearly, called a dataset that is linearly separable [25]. If the dataset is not linearly separable, then the kernel trick is used, namely changing the dimensions of the feature space so that a hyperplane can be created that divides the [26] dataset in half.

Furthermore, MLP is a type of ANN that has at least three layers of neurons: input, hidden, and output [27]. MLP goes through a learning process where, in each iteration of the learning process, the weights and biases of each neuron are adjusted to minimize loss to actual labels [28]. Logistic regression transforms the features so that each feature with a linear relationship with its label is pinned to a binary label [29].

3.2. ROC Threshold Selection

An imbalanced dataset is when, in a dataset, the number of one label (majority label) is far greater than the other (minority labels) [30]. Imbalanced datasets can affect the performance of machine learning models, then the validity of a measurement metric [31]. Metric accuracy, precision, and f1-score do not reflect the ability of a machine learning model if the dataset is imbalanced. On the other hand, studies used sensitivity and specificity in imbalanced datasets with binary labels because these metrics focus on the predictive ability of each label without being mixed with other labels [32]. Besides that, the g-mean is also a good metric because the g-mean aggregates sensitivity and specificity values.

ROC is also a measurement metric that can be used in imbalanced datasets because ROC is a curve that shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) when the threshold in a machine learning model is changed [33]. TPR is another name for sensitivity and FPR = 1 - specificity. AUC is a quantitative measure of an ROC. The larger the AUC value, the better [34]. The algorithm in Figure 2 shows the optimal threshold selection algorithm. Machine learning methods generally choose a threshold that maximizes accuracy. However, accuracy does not reflect its actual performance in imbalanced datasets. The optimum threshold can be determined from the ROC curve. One point on the ROC curve reflects the TPR and FPR at a threshold. If we calculate the G-mean of each threshold, then the most optimal threshold is the one with the largest G-mean value.

Algorithm 1: Optimal Threshold Selection Algorithm

Data: $Soft\_Labels, Actual\_Labels$
Result: $Optimal\_Theshold$
$Thresholds, Sensitivities, Specificities \leftarrow ROC\_Process(Soft\_Labels, Actual\_Labels)$;
for $(Sensitivity \in Sensitivities) \cup (Specificity \in Specificities)$ **do**
  Calculate G-Mean with the following formula:

$$G\_Mean = \sqrt{Sensitivity \times Specificity} \quad (1)$$

end
$Optimal\_Theshold \leftarrow Thresholds[ArgMax(G\_Means)]$;

Figure 2. The optimal threshold selection algorithm

3.3. GA feature selection for AUC Maximization

GA is an optimization algorithm inspired by natural selection and genetics [35]. Finding the optimum value in GA is an iterative process in which, at each iteration, a population with a certain number is generated. The population is the composition of the previous population (parent), mutation, and crossover. Each population is the input of an objective function, where in each generation, each population executes that objective function. The population selection in the next iteration is based on the results of executing the objective function. Our GA uses five iterations. The population size is 25. Mutation, crossover, and parent probabilities are all 0.3, 0.5, and 0.1, respectively.

Because feature selection is an optimization method for a prediction model, GA can also be used for feature selection [36]. In our feature selection, one individual in the population represents an integer value from the Boolean filter of the original features. The upper boundary of the individual is calculated with the following equation:

$$\text{Upper\_Boundary} = 2^{F_o} \quad (2)$$

where F_o is the number of original features.

After doing GA parameter settings, GA runs where each individual operates the objective function in each iteration. Usually, the objective function in feature selection is accuracy. Still, in the case of imbalanced datasets, we propose a novel GA feature selection with the objective function for AUC maximization, ImbGAFS. The algorithm in Figure 3 shows the algorithm of the objective function. Input has a range from 0 to *Upper_Boundary*. Some of the uniqueness of ImbGAFS compared to other GA feature selections is the use of optimal threshold selection and G-Mean as the final metric of the objective function. Note that the ImbGAFS still adopts the SMOTE process.

Algorithm 2: ImbGAFS Objective Function Algorithm

Data: $Input, Original\_Feature$
Result: G_Mean
$Binary\_Input \leftarrow Convert\_to\_Binary(Input)$;
$Boolean\_Filter \leftarrow Convert\_to\_Array(Binary\_Input)$;
$Sub\_Feature \leftarrow Original\_Feature[Boolean\_Filter]$;
$ML\_Model \leftarrow Train\_Model(Sub\_Feature)$; /* (with Algorithm 1) */
$G\_Mean \leftarrow Calculate\_G\_Mean(ML\_Model)$;

Figure 3. The ImbGAFS objective function algorithm

4. RESULTS AND DISCUSSION

4.1. Results

We use 6,561 data items from the dataset. In the pre-processing stage, we found 6,512 data items in the dataset that had missing values. We fill in the missing value with zero imputation. We run a Pearson correlation analysis on the features in the dataset. Figure 4 shows a heatmap of its Pearson correlation matrix. The most important part of this matrix is the bottom row, namely the correlation between features and labels. The feature with the strongest correlation is 'Effect: Impact to flight' with a value of -0.21. The features with the weakest correlation were 'Record ID' and 'Wildlife: Species.' Both of which have a Pearson correlation value of -0.01.

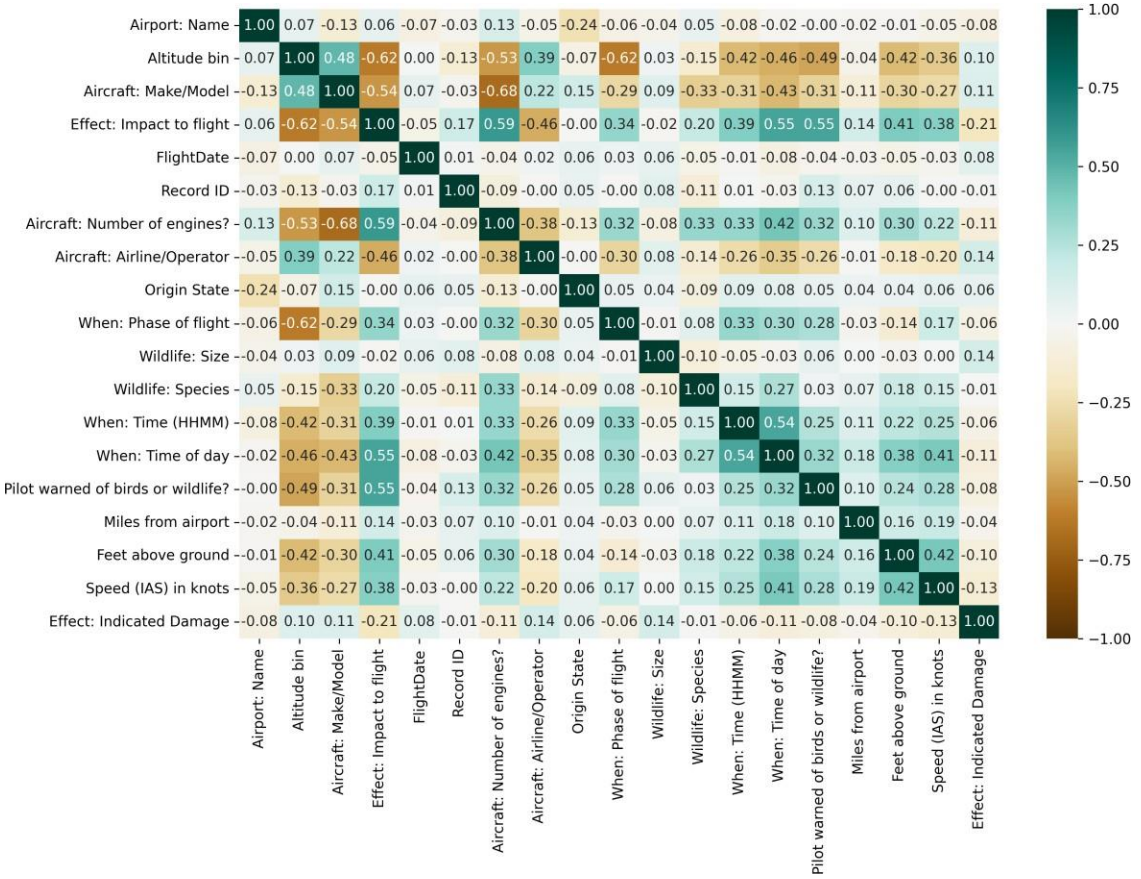
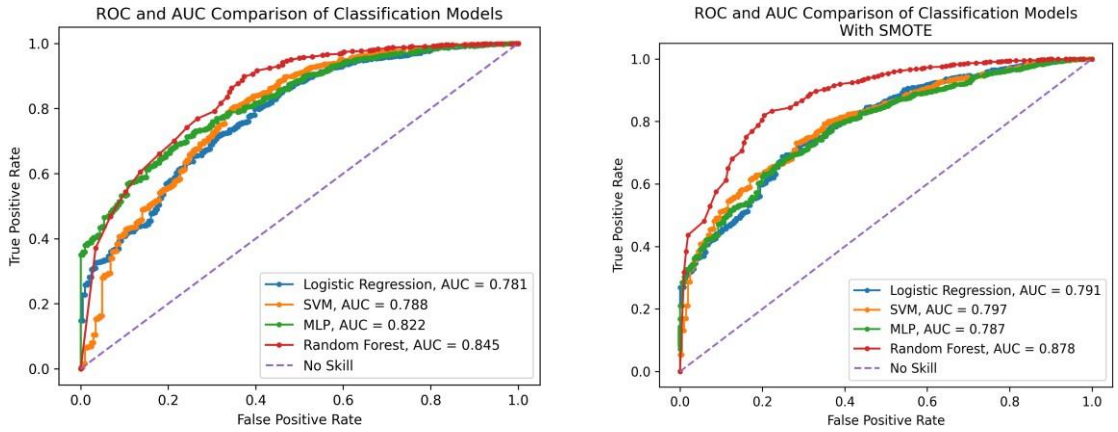


Figure 4. Pearson correlation analysis on bird strike dataset features

We continue pre-processing by standardizing the features. Before standardization, the average feature value is 308.1, while the standard deviation is 908.4. After standardization, the feature mean is 13 x 10^{-18}, while the standard deviation is 1. The goal of standardization is for the mean to be 0 and the standard deviation to be 1. We run the label encoder. After the label encoder, the "Caused damage" data items become 0, while "No damage" becomes 1. There are 411 data items with label 0, while 6,150 data items have the label 1. Label 0 is the minority label with a proportion of 6.3% of the dataset. This figure categorizes the dataset into an imbalanced dataset with a moderate level.

We divide the dataset into training data and testing data. 50% of the dataset is for training data, while the other 50% of the dataset is for testing data. We run SMOTE on the training data as an oversampling method. After SMOTE, the number of training datasets with labels 0 and 1 is 3,075. The next step is machine learning model training. We have four models to compare: random forest, SVM, logistic regression, and MLP. We compare each model with and without SMOTE. So, in total, we compared eight models. Figure 5 shows the ROC comparison of the eight models. SMOTE improved the performance of three models: logistic regression, SVM, and random forest. In MLP, SMOTE lowers the model's performance from AUC = 0.822 to AUC = 0.787. The random forest is the model with the best performance, be it without SMOTE or with SMOTE. The highest performance belongs to random forest + SMOTE, with a value of 0.878. At the same time, the lowest performance belongs to logistic regression without SMOTE, with a value of 0.781.



(a) (b)

Figure 5. ROC comparison of four machine learning methods: logistic regression, SVM, MLP, and random forest: (a) Without SMOTE (b) With SMOTE

We adopt the random forest model with the best performance for the next step. In the next step, we use optimal threshold selection based on the ROC curve from the random forest. We apply the algorithm in Figure 2 to the random forest model. Figure 6 is a bar chart that shows the results. Sensitivity, which highlights the model's ability to predict the majority label, performs better than after threshold selection. But with a contrast specificity score on the minority label, the g-mean result is 0.5199. Although sensitivity decreased, specificity increased by 2.9x. With this increase, the random forest g-mean after optimal threshold selection is 0.8037. Optimal threshold selection can increase the g-mean by 1.6x.

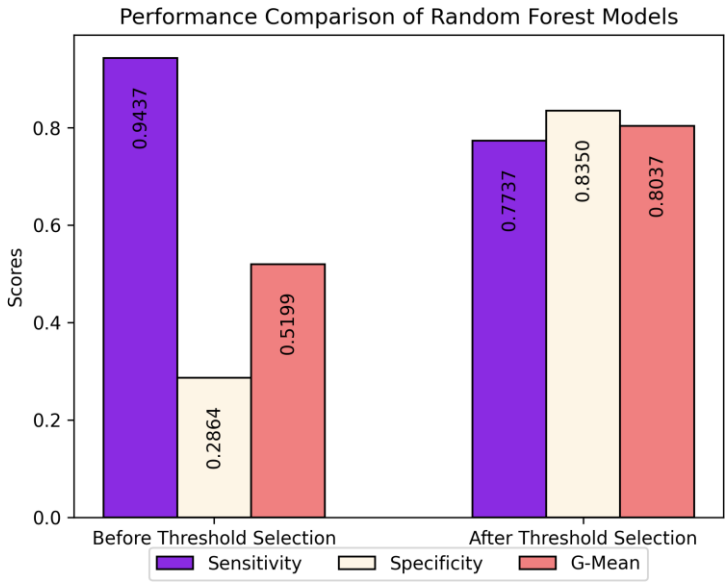


Figure 6. The influence of optimal threshold selection on sensitivity, specificity, and g-mean

In the following test, we operate ImbGAFS. The first step is to carry out an objective function landscape analysis. Figure 7 shows the objective function landscape in our case study. The first aspect observed is the number of valleys in the landscape. Observation result shows many valleys, which indicates that there may be a trap in the local optima. The large number of valleys can also influence the choice of population size. A large population size amidst many valleys can help with wider exploration. On the influence of exploration strategies, many valleys indicate a complex problem. A large value of crossover probability can promote exploration.

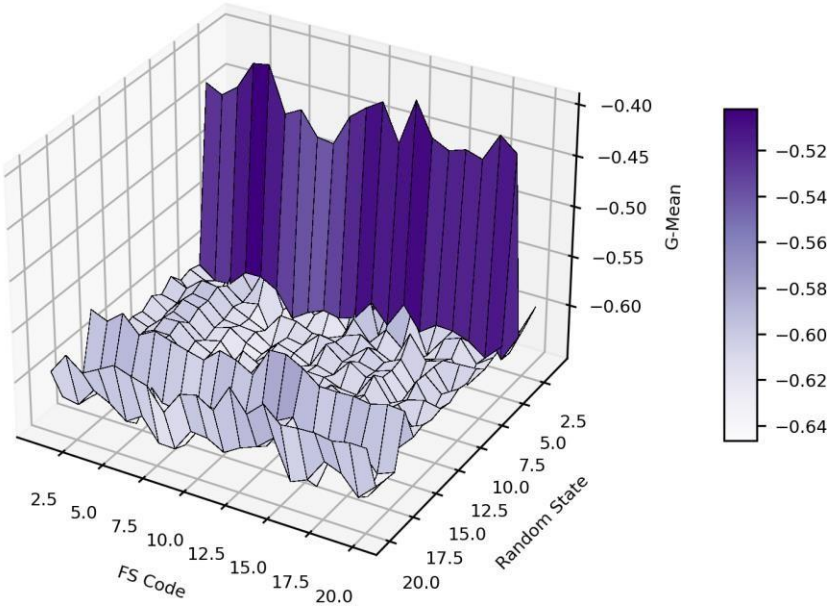


Figure 7. The objective function landscape of our ImbGAFS method

We ran ImbGAFS with five iterations and 25 populations. Figure 8 shows the fitness curve. The first aspect that can be observed is fitness improvement. Since this is a maximization problem, increasingly better g-mean values are a good sign. Then, in terms of convergence, since the 4th generation, the value has not decreased, which shows that the optimum value has been obtained. Finally, the sharp increase in the middle generation shows that ImbGAFS is exploring.

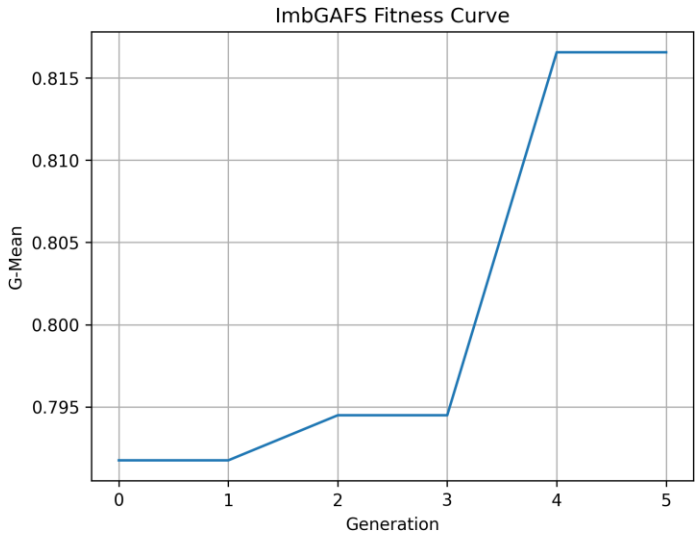


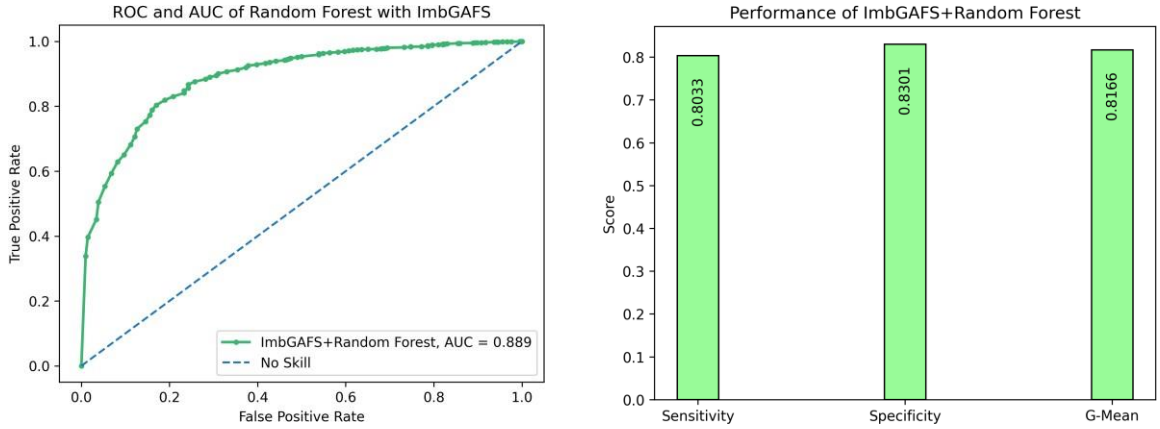
Figure 8. The fitness curve of the ImbGAFS execution

Table 2 shows the result of feature selection from ImbGAFS. We also present the Pearson correlation score for comparison. We sort features based on their Pearson correlation values. Eliminated features with the best Pearson correlation scores show that ImbGAFS has a stochastic method. The order of a value does not affect the generation of its value.

Table 2. The Feature Selection Result of ImbGAFS

Feature Name	Pearson Correlation Score	ImbGAFS Result
'Effect: Impact to flight'	0.20660458	Selected
'Aircraft: Airline/Operator'	0.14475705	Eliminated
'Wildlife: Size'	0.14415801	Selected
'Speed (IAS) in knots'	0.12556103	Eliminated
'Aircraft: Number of engines?'	0.1144976	Selected
'When: Time of day'	0.10675161	Selected
'Aircraft: Make/Model'	0.10560883	Selected
'Altitude bin'	0.09788964	Selected
'Feet above ground'	0.09759895	Selected
'FlightDate'	0.08181713	Eliminated
'Pilot warned of birds or wildlife?'	0.07988008	Selected
'Airport: Name'	0.07580298	Selected
'When: Phase of flight'	0.06384834	Eliminated
'Origin State'	0.06016311	Selected
'When: Time (HHMM)'	0.05944301	Eliminated
'Miles from airport'	0.03624733	Selected
'Record ID'	0.00768932	Selected
'Wildlife: Species'	0.00507544	Selected

Finally, we show the random forest model results from ImbGAFS. Figure 5 shows the performance of the model. The first is the ROC curve. Figure 5(a) shows that the ROC from random forest results from ImbGAFS is better than the conventional method without feature selection. The increase in the AUC value is from 0.878 to 0.889. Then, after carrying out optimal threshold selection, ImbGAFS+random forest also has better sensitivity, specificity, and g-mean than conventional methods. The increase is from 0.7737, 0.8350, and 0.8037 to 0.8033, 0.8301, and 0.8166, respectively.



(a) (b)

Figure 5. The ImbGAFS+Random Forest Performances: (a) The ROC Curve (b) The threshold selection-based sensitivity, specificity, and g-mean.

4.2. Discussion

Several studies discuss airplane failure prediction due to bird strikes, such as papers [9], [10]. However, these studies need to analyze the imbalance in the dataset further. We see that the minority labels in our dataset have a proportion of 6.3% compared to the entire data set, while the category is moderately imbalanced. Our research contribution is an airplane failure prediction by bird strike, which applies various methods to deal with imbalanced datasets.

One of the optimization methods in imbalanced datasets is ROC threshold selection, as used by previous papers in the field of lithium-ion batteries [11] and diabetes mellitus [12]. On the other hand, AUC maximization is a method other than ROC threshold selection to improve prediction performance on imbalanced datasets, which has been applied with stochastic gradient [13] and PSVM [14]. Using AUC maximization with ROC threshold selection is a research opportunity. Our research contributes to using a combination of AUC maximization and ROC threshold selection.

Finally, the GA method is one of the most superior methods in feature selection. The use of GA in feature selection in the research of Alawad *et al.* [15] can improve the performance of extra tree classifier, random forest, support vector machine, and KNN in identifying brain haemorrhage. In research by Yang *et al.* [16], GA is used in conjunction with a time-series feature extractor (Tsfresh) to extract signals and select the best features from IoT data streams for anomaly detection. The classification uses XGBoost. Using GA feature selection with g-mean threshold selection as an objective function is a research opportunity. Our research contribution is a novel GA feature selection method called ImbGAFS, which can be used as feature selection to improve model performance on imbalanced datasets.

5. CONCLUSIONS

We succeeded in implementing a novel method called ImbGAFS, namely GA feature selection for imbalanced datasets. We got an imbalanced dataset from the bird strike dataset by the FAA, where the imbalance dataset category is moderate, with minority labels accounting for 6.3% of the entire dataset. Our test results show that random forest is the best machine learning method in airplane failure prediction compared to SVM, logistic regression, and MLP. SMOTE can increase random forest AUC from 0.845 to 0.878. Finally, the random forest model from ImbGAFS is better than the conventional method without feature selection. The increase in the AUC value is from 0.878 to 0.889. Then, after carrying out optimal threshold selection, ImbGAFS+random forest also has better sensitivity, specificity, and g-mean than conventional methods. The increase is from 0.7737, 0.8350, and 0.8037 to 0.8033, 0.8301, and 0.8166, respectively. Future work can direct proof of the application of ImbGAFS to imbalance problems in other datasets.

ACKNOWLEDGEMENTS

We would like to thank the Human Resource (SDM) Department of Telkom University for funding the accommodation of this research project. We would also like to thank the research and community service (PPM) Department of Telkom University for funding this publication through the Doctoral Dissertation Research (PDD) funding scheme.

REFERENCES

[1] C.-K. Chen, W.-T. Juan, Y.-C. Liang, P. Wu, and C.-M. Chuong, “Making region-specific integumentary organs in birds: evolution and modifications,” Curr. Opin. Genet. Dev., vol. 69, pp. 103–111, 2021.

[2] T. Sustainability, “Stories of Sustainability: Birds, Bees, and Trees,” Clim. Change Sustain. Environ. Justice, 2020.

[3] A. S. Ramadhan, M. Abdurohman, and A. G. Putrada, “WSN based agricultural bird pest control with buzzer and a mesh network,” in 2020 8th International Conference on Information and Communication Technology (ICoICT), IEEE, 2020, pp. 1–5.

[4] K. Celikmih, O. Inan, and H. Uguz, “Failure prediction of aircraft equipment using machine learning with a hybrid data preparation method,” Sci. Program., vol. 2020, pp. 1–10, 2020.

[5] F. Lu, J. Wu, J. Huang, and X. Qiu, “Aircraft engine degradation prognostics based on logistic regression and novel OS-ELM algorithm,” Aerosp. Sci. Technol., vol. 84, pp. 661–671, 2019.

[6] W. Yan, “Application of random forest to aircraft engine fault diagnosis,” in The Proceedings of the Multiconference on" Computational Engineering in Systems Applications", IEEE, 2006, pp. 468–475.

[7] R. N. Toma, A. E. Prosvirin, and J.-M. Kim, “Bearing fault diagnosis of induction motors using a genetic algorithm and machine learning classifiers,” Sensors, vol. 20, no. 7, p. 1884, 2020.

[8] K. T. Chui, B. B. Gupta, and P. Vasant, “A genetic algorithm optimized RNN-LSTM model for remaining useful life prediction of turbofan engine,” Electronics, vol. 10, no. 3, p. 285, 2021.

[9] S. Nimmagadda, S. Sivakumar, N. Kumar, and D. Haritha, “Predicting airline crash due to birds strike using machine learning,” in 2020 7th international conference on smart structures and systems (ICSSS), IEEE, 2020, pp. 1–4.

[10] S. Misra, I. Toppo, and F. A. C. Mendonca, “Assessment of aircraft damage due to bird strikes: a machine learning approach,” Int. J. Sustain. Aviat., vol. 8, no. 2, pp. 136–151, 2022.

[11] O. Ojo, H. Lang, Y. Kim, X. Hu, B. Mu, and X. Lin, “A neural network based method for thermal fault detection in lithium-ion batteries,” IEEE Trans. Ind. Electron., vol. 68, no. 5, pp. 4068–4078, 2020.

[12] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, “Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods,” BMC Med. Inform. Decis. Mak., vol. 22, no. 1, p. 36, 2022.

[13] Y. Yan, Y. Xu, Q. Lin, L. Zhang, and T. Yang, "Stochastic Primal-Dual Algorithms with Faster Convergence than O(1/√T) for Problems without Bilinear Structure," ArXiv Prepr. ArXiv190410112, 2019.
[14] G. Wang, S. W. H. Kwok, M. Yousufuddin, and F. Sohel, "A Novel AUC Maximization Imbalanced Learning Approach for Predicting Composite Outcomes in COVID-19 Hospitalized Patients," IEEE J. Biomed. Health Inform., 2023.
[15] D. M. Alawad, A. Mishra, and M. T. Hoque, "AIBH: accurate identification of brain hemorrhage using genetic algorithm based feature selection and stacking," Mach. Learn. Knowl. Extr., vol. 2, no. 2, pp. 56–77, 2020.
[16] Z. Yang, I. A. Abbasi, E. E. Mustafa, S. Ali, and M. Zhang, "An anomaly detection algorithm selection service for IoT stream data based on tsfresh tool and genetic algorithm," Secur. Commun. Netw., vol. 2021, pp. 1–10, 2021.
[17] A. Anggraini and S. Titis Setyabudi, "Heroism In Clint Eastwood's Miracle On The Hudson Movie (2016): A Psychological Approach," PhD Thesis, Universitas Muhammadiyah Surakarta, 2021.
[18] R. A. Dolbeer, M. J. Begier, P. R. Miller, J. R. Weller, A. L. Anderson, and others, "Wildlife Strikes to Civil Aircraft in the United States, 1990–2019," United States. Department of Transportation. Federal Aviation Administration …, 2021.
[19] I. D. Oktaviani and A. G. Putrada, "KNN imputation to missing values of regression-based rain duration prediction on BMKG data," J. Infotel, vol. 14, no. 4, pp. 249–254, 2022.
[20] T. Purwoningsih, H. B. Santoso, K. A. Puspitasari, and Z. A. Hasibuan, "Early Prediction of Students' Academic Achievement: Categorical Data from Fully Online Learning on Machine-Learning Classification Algorithms," J. Hunan Univ. Nat. Sci., vol. 48, no. 9, 2021.
[21] M. B. Satrio, A. G. Putrada, and M. Abdurohman, "Evaluation of Face Detection and Recognition Methods in Smart Mirror Implementation," in Proceedings of Sixth International Congress on Information and Communication Technology, Springer, 2022, pp. 449–457. doi: https://doi.org/10.1007/978-981-16-2380-6_39.
[22] A. G. Putrada, N. Alamsyah, M. N. Fauzan, and S. F. Pane, "NS-SVM: Bolstering Chicken Egg Harvesting Prediction with Normalization and Standardization," JUITA J. Inform., vol. 11, no. 1, pp. 11–18, 2023.
[23] A. G. Putrada and D. Perdana, "Improving Thermal Camera Performance in Fever Detection during COVID-19 Protocol with Random Forest Classification," in 2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), IEEE, 2021, pp. 1–6.
[24] M. Ameliasari, A. G. Putrada, and R. R. Pahlevi, "An Evaluation of SVM in Hand Gesture Detection Using IMU-Based Smartwatches for Smart Lighting Control," J. INFOTEL, vol. 13, no. 2, pp. 47–53, 2021.
[25] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, "CIMA: A Novel Classification-Integrated Moving Average Model for Smart Lighting Intelligent Control Based on Human Presence," Complexity, vol. 2022, no. Article ID 4989344, p. 19, 2022.
[26] B. A. Fadillah, A. G. Putrada, and M. Abdurohman, "A Wearable Device for Enhancing Basketball Shooting Correctness with MPU6050 Sensors and Support Vector Machine Classification," Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control, 2022.
[27] R. R. Pamungkas, A. G. Putrada, and M. Abdurohman, "Performance Improvement of Non Invasive Blood Glucose Measuring System With Near Infra Red Using Artificial Neural Networks," Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control, vol. 4, no. 4, 2019.
[28] M. F. Akbar, A. G. Putrada, and M. Abdurohman, "Smart Light Recommending System Using Artificial Neural Network Algorithm," in 2019 7th International Conference on Information and Communication Technology (ICoICT), IEEE, 2019, pp. 1–5.
[29] D. C. E. Saputra, Y. Maulana, E. Faristasari, A. Ma'arif, and I. Suwarno, "Machine Learning Performance Analysis for Classification of Medical Specialties," in Proceeding of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics: ICEBEHI 2022, 5–6 October, Surabaya, Indonesia, Springer, 2023, pp. 513–528.
[30] A. G. Putrada, I. D. Wijaya, and D. Oktaria, "Overcoming Data Imbalance Problems in Sexual Harassment Classification with SMOTE," Int. J. Inf. Commun. Technol. IJoICT, vol. 8, no. 1, pp. 20–29, 2022.

[31] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, "XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data," in 2022 International Symposium on Electronics and Smart Devices (ISESD), IEEE, 2022, pp. 1–7.
[32] Z. Xu, D. Shen, Y. Kou, and T. Nie, "A synthetic minority oversampling technique based on Gaussian mixture model filtering for imbalanced data classification," IEEE Trans. Neural Netw. Learn. Syst., 2022.
[33] J. Singla and others, "Comparing ROC curve based thresholding methods in online transactions fraud detection system using deep learning," in 2021 international conference on computing, communication, and intelligent systems (ICCCIS), IEEE, 2021, pp. 9–12.
[34] A. G. Putrada and M. Abdurohman, "Anomaly Detection on an IoT-Based Vaccine Storage Refrigerator Temperature Monitoring System," in 2021 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), IEEE, 2021, pp. 75–80.
[35] T. Alam, S. Qamar, A. Dixit, and M. Benaida, "Genetic algorithm: Reviews, implementations, and applications," ArXiv Prepr. ArXiv200712673, 2020.
[36] W. Ali and A. A. Ahmed, "Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting," IET Inf. Secur., vol. 13, no. 6, pp. 659–669, 2019.

AUTHORS

A.G. Putrada received an M.Tech from ITB, Bandung, Indonesia, and did his Bachelor's degree in electrical engineering. He is pursuing his Ph.D. in Computer Science from Telkom University. His research interests include smart lighting, machine learning, and edge computing.Short Biography

S. Prabowo received a bachelor's and master's degree in informatics Engineering from Telkom University in 2011 and 2014, respectively. Currently, he is pursuing a doctoral degree at Telkom University, Bandung. His dissertation concerns the security framework for the Internet of things in Indonesia.