



Firefly Algorithm Under-sampling for Imbalance Data in Breast Cancer Survival Prediction

Diya Namira Purba¹, Namora Purba²

¹Doctoral Degree of Computer Science, Department of Mathematics and Natural Sciences, Gadjah Mada University, Yogyakarta, Indonesia

²Chatib Quzwain Regional General Hospital, Sarolangun, Jambi, Indonesia

¹diyanamirapurba@mail.ugm.ac.id, ²clinicnamora@gmail.com

Abstract

Breast cancer remains a major health challenge, affecting approximately 1.7 million individuals annually and often leading to severe complications. Predicting survival outcomes is difficult due to highly imbalanced data, with 3,408 death cases compared to only 616 survival cases. To address this issue, we applied the Firefly Algorithm-based under-sampling (FAUS) to balance the dataset and combined it with three machine learning classifiers: Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN). Experimental results show that FAUS substantially improves predictive performance compared to conventional under-sampling. Among the tested models, RF achieved the highest F1-score of 0.79, while DT and KNN reached 0.72 and 0.68, respectively. The results indicate that FAUS is effective in preserving representative samples, thereby enhancing model performance in breast cancer survival prediction.

Keywords: breast cancer; disease prediction; firefly algorithm; imbalance data; machine learning

How to Cite: D. N. Purba and Namora Purba, "Firefly Algorithm Under-sampling for Imbalance Data in Breast Cancer Survival Prediction", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 1, pp. 127 - 132, Feb. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i1.6439>

Received: March 2, 2025

Accepted: October 18, 2025

Available Online: February 25, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Cancer is considered a fatal disease due to its high mortality rate [1]. One of the cancers that frequently happen among women is breast cancer [2]. According to the Agency for Research, there will be 1.7 million breast cancer cases in 2021[2]. This cancer can spread to other parts of the body parts, and it is also a type of metastatic cancer [2]. The presence of a lump in the breast is the most frequent early sign of breast cancer, with a total of 75% of cases [2]. The symptoms of breast cancer have significantly impacted the patient's physical and well-being [1]. This symptom can lead to complication, including cancer progression and metastasis [3]. Predicting breast cancer is crucial in helping clinicians assess the patient's prognosis, determine the treatment plans, and improve overall patient outcomes [1].

Recently, clinicians have used a basic statistical approach to identify new variables from the data. Previously, the clinicians used basic analytical software such as Microsoft Excel, SPSS, and STATA. This approach is not adaptable to identifying new variables

to predict survival factors in the dataset [2]. To handle this issue, machine learning is one of the computational methods that can be used as an alternative method to predict the survival of breast cancer cases [3]. This method can help identify the new variables affecting the survival cases of breast cancer [4]. One of the challenges to predicting survival cases is the significant difference between survival and death cases, resulting in an imbalance of data that can reduce the prediction performance [5].

To handle this imbalance issue, various sampling method can be implemented to balance the target features [6]. There are three sampling method i.e. under-sampling, oversampling and hybrid-sampling. The under-sampling is a method that reduce the majority class until its balanced with the minority classes [7]. Oversampling is a method that create a new instances from the minority class until it has the same total with the majority class [8]. Lastly, hybrid-sampling is a combination of under-sampling and oversampling [9]. Despite the advantages of oversampling and hybrid sampling, under sampling was

chosen in this study due to its ability to handle large imbalanced datasets.

Several studies have explored breast cancer survival prediction using machine learning and various sampling techniques. In 2022, Aditi et al. performed a study about breast cancer survival prediction from the imbalanced dataset with machine learning prediction [1]. This study used a breast cancer dataset from SEER [1]. The sampling methods used in this dataset are under-sampling and oversampling with the machine learning algorithms, which are Decision Tree (DT), Random Forest (RF), K Nearest Neighbor (KNN), and Support [7] Vector Machine (SVM) [1]. This study shows that SVM achieved the best model with a value of 1 for precision and 0.99 for the Area Under the Curve (AUC) [1]. Another study about cancer detection using machine learning and resampling techniques was proposed by Gurcan et al. in 2024 [4]. This study used a breast cancer dataset from SEER [4]. The machine learning models are Multi-Layer Perceptron (MLP), RF, Logistic Regression (LR), Deep Neural, and XGBoost with SMOTEENN [4]. In this model, RF achieved the best performance with the value 94.69 % with the value of mean score [4].

Meanwhile, a study about an implementation about voting-based machine learning technique to detect breast cancer with feature selection and SMOTE-ENN class balancing proposed by Chhiillar in 2024 [10]. The soft voting-based machine learning models utilized in this model are MLP, Support Vector Machine (SVM), and XGBoost and SMOTE-ENN for the sampling method [10]. The soft voting-based machine learning models achieved high scores with an accuracy 99.42% [10].

Another related study about imbalance data cancer prediction with the implementation of a machine learning model-based proposed by Tran et al. in 2022[6]. This study used a Wisconsin breast cancer dataset [6]. The engineered up-sampling method (ENUS) with Decision Tree (DT), RF, Neural Networks (Nnet), MLP, Extra Gradient Boost Tree (XGBTree) and SVM for the machine learning models [6]. This study shows that ENUS achieved the best performance with the value of balanced accuracy of 97% [6].

In 2024, a study about firefly algorithm under-sampling in handle imbalance data in DTI prediction [11]. This study used a gold standard dataset. The study implements the FAUS with Random Forest Algorithm [11]. From the result, this study achieved with the value of accuracy 0.94 [11].

Previous studies have identified various sampling methods, such as under-sampling, over-sampling, and hybrid-sampling methods. Oversampling techniques include SMOTE and its variations (SMOTE-ENN, SMOTE-Tomek, ADASYN) [8]. However, these methods may introduce synthetic bias and increase computational complexity. The hybrid sampling method combines over-sampling and under-sampling.

These methods require additional parameter tuning, making them computationally expensive [7].

Despite these advances, most existing works emphasize the novelty of applications rather than methodological innovations in sampling. Furthermore, while Firefly Algorithm-based under-sampling (FAUS) has been applied in DTI prediction, its methodological advantage in medical survival prediction tasks has not been thoroughly investigated.

Thus, this study addresses the gap by evaluating FAUS as an optimization under-sampling for breast cancer survival prediction. The focus is not only on its application to a new dataset but also on its methodological potential to preserve representative samples and improve predictive performance compared to conventional resampling methods.

2. Methods

The proposed method begins with the SEER breast cancer survival dataset, which exhibits a severe class imbalance between death and survival cases. To address this issue, the FAUS is applied to generate optimized subsets of the majority class. FAUS leverages the swarm-intelligence behavior of fireflies to identify and retain informative instances while discarding redundant or unrepresentative samples.

2.1 Dataset

We used a SEER Breast Cancer dataset [12]. This dataset has two classes, i.e. survival and death cases. There are 3404 death cases and 616 survival cases obtained from the dataset [2], which presented in Figure 1. This dataset can be retrieved from this link <https://www.kaggle.com/datasets/sujithmandala/seer-breast-cancer-data>. The dataset contains several factors influencing the prognosis and survival of breast cancer patients such as patient demographics, tumor characteristics, and treatment modalities[12]. The target column for this dataset is survival outcome [12]. This dataset is split into 70% of train dataset and 30% of test dataset.

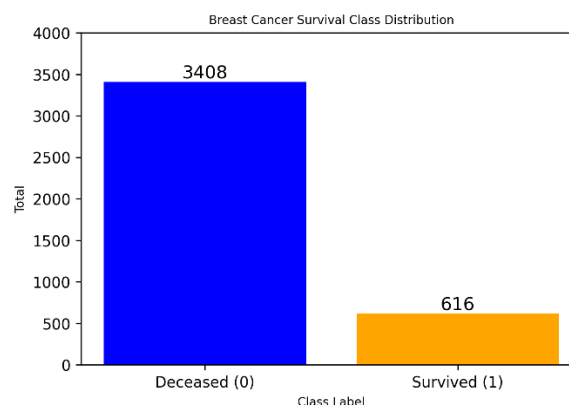


Figure 1. The Label Data Distribution

The distribution of cases number of each class in train and test set is shown in Figure 2. In the train datasets,

the number of death cases is 2385 cases with 431 of survival cases. The difference of cases is significant, making the distribution imbalance. For test datasets, the number of death cases are 1023 death cases with 185 survival cases. Therefore, the distribution of class samples in both train and test datasets are imbalanced.

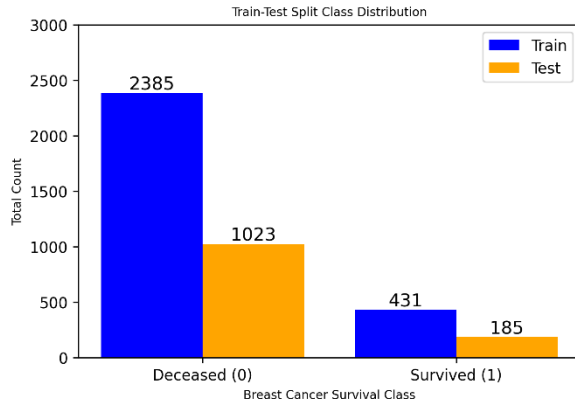


Figure 1. The Label Distribution on Train and Test Data

2.2 Firefly Algorithm

The Firefly Algorithm (FA) is one of the metaheuristic algorithms introduced by Yang in 2009 [11]. The algorithm was inspired by the natural behavior of the fireflies [11]. The brightness of the fireflies is used to attract the other fireflies. Each fireflies have different level of brightness [11]. The brightest fireflies will attract the less bright one [11].

The algorithm can be combined with the under-sampling-based algorithm [11]. In the first process, the algorithm initializes the fireflies with a random position representing a total of the majority class [13]. The intensity calculation determines the attractiveness of each firefly [13]. Brighter fireflies attract each other, leading to position updates [5].

The iterative process will lead the fireflies to update their position until they reach the maximum iteration [14]. To evaluate the intensity of each firefly will be based on the result from the F1-score calculation. This intensity represents the brightness of the social behavior of the fireflies [14]

After all processes reach the maximum iteration, the algorithm will result in the top three best fireflies. The FAUS flowchart is presented in Figure 3. The F1 score is an evaluation metric based on the harmonic mean of precision and recall [15]. Precision is the number of predicted positive instances from all instances classified as positive by the model [10]. Recall is the number of actual positive instances correctly predicted by the model [16]. The formula for firefly intensity is formulated in Equation 1.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1)$$

The parameters that used in Firefly Algorithm Under-sampling aimed to explore the search process of the fireflies to find the best solution [17]. The primary

reference about the algorithm process is a book that introduces the Firefly algorithm by Yang [18]. The selection of hyperparameter for FAUS was based on previous research [11]. Parameter ranges of the algorithm is presented in Table 1.

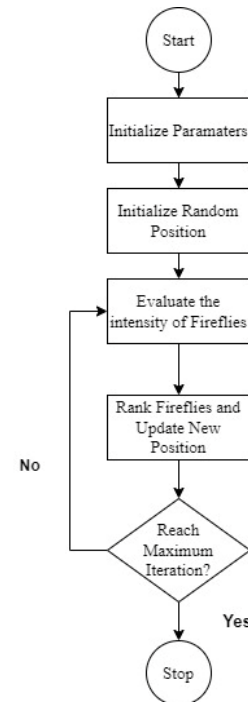


Figure 2. Firefly Algorithm Under-sampling Process [11]

Regarding parameterization, we employed 10 fireflies and 10 iterations, which may appear relatively small compared to typical metaheuristic configurations. However, preliminary experiments indicated that increasing the population size or the number of iterations did not produce a notable improvement. These findings suggest that a configuration of 10 fireflies is sufficient to achieve convergence and to identify optimal representative samples within this dataset.

Table 1. Firefly Algorithm Parameter Ranges

Parameters Name	Range of Value
Number of fireflies	10
Maximum Iterations	10
Alpha	1
Betha	0.8
Gamma	0.9

Additionally, a limited sensitivity analysis was conducted to assess robustness. The results demonstrated that performance trends remained consistent across different parameter settings, confirming that using 10 fireflies yields stable and reproducible outcomes without performance degradation [19].

2.3 Under-sampling Method

Under-sampling is a widely used technique for addressing class imbalance in classification problems, particularly when the majority class significantly outnumbers the minority class. This method works by

reducing the number of majority class instances, thereby balancing the class distribution and mitigating bias in the learning algorithm [19]. While under sampling can improve the classifier performance by preventing the model from being overwhelmed by the majority class, it also carries the risk of discarding potentially valuable information. To overcome this limitation, more advanced strategies such as heuristic based under sampling have been developed to retain the most informative majority class samples selectively [9].

The natural flashing behavior of fireflies inspires the FAUS method, which offers a metaheuristic approach to under-sampling by selectively retraining the most informative majority class samples. In this method, each firefly represents a candidate subset of the majority data, and its brightness is measured using a fitness function, often linked to classification performance. Less bright fireflies move toward brighter ones, allowing better solutions to guide the search process.

Through iterative updates, the algorithm converges on an optimal subset of majority instances in the dataset without discarding crucial information. By combining exploration and exploitation, FA-based under-sampling effectively reduces class imbalance, minimizes bias, and improves the robustness of classification models compared to other sampling methods [20].

2.4 Prediction Model

In this study, the prediction models were developed by implementing three algorithms, i.e., Random Forest, Decision Tree, and K nearest Neighbor. Random Forest is a classification algorithm based on multiple decision trees during the training and test process [21]. Moreover, this algorithm can construct multiple decision trees to avoid overfitting from train data [21]. Recently, this algorithm can be implemented on categorical features, measure the feature importance, outlier detection, etc [22].

Decision tree is a supervised learning algorithm [23]. This algorithm implements a recursive partitioning approach for the model decision process [23]. The algorithm has several criteria, such as Gini, Impurity, Entropy, and Gain ratio [23]. Moreover, this algorithm can optimize the classification or regression accuracy [23]. K-nearest neighbor (KNN) is one of the classification algorithms based on the k value [24]. This algorithm processes the distance between the instance with another instance [24]. The distance between the instances is calculated by Euclidean distance, Manhattan distance, and Minkowski distance [24].

2.5 Model Evaluation Metric

After the prediction models were developed, the confusion matrix is a performance evaluation metric used in classification [15]. This metric is represented in Table 2.

Accuracy is an evaluation metric that measures predictions for positive and negative instances[23].

Precision is the true positive rate divided by the addition of true positive and false positive, measuring the accuracy of positive predictions [25]. Recall, also known as sensitivity, is the ratio of true positive rate divided by the addition of true positive rate and false negatives, indicating the model's ability to correctly identify positive instances [22]. Lastly, the F1-score is an evaluation metric that measures precision and recall [15]. The formula of accuracy, precision, and recall are presented in Equations 2-4.

Table 2. Confusion Matrix

Predicted Values	Actual Values		
		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3. Results and Discussions

The following section presents the results and discussion of the experiments conducted in this study to evaluate the effectiveness of the FAUS in addressing class imbalance within the SEER breast cancer survival dataset.

3.1 Model Development

Hyperparameter tuning was implemented to determine the best parameters for all models. The tuning process aimed to enhance the classification model performance. The following are the default parameters used in hyperparameter tuning, which are set according to the standard rules of the algorithm, which presented in Table 3.

Table 3. Hyperparameter Tuning Ranges

Method	Parameters	Ranges
DT	Criterion	['gini', 'entropy']
	Max_depth	[3, 5, 10]
	Min_samples_split	[2, 5, 10]
	Min_samples_leaf	[1, 2, 4]
	Max_features	['sqrt', 'log2']
RF	N_estimators	[50, 100, 200, 500]
	Max_depth	[10, 20, 30]
	Min_samples_leaf	[1, 2, 4]
	Min_samples_split	[2, 5, 10]
KNN	N_neighbors	[3, 5, 7, 9, 11]
	Weights	['uniform', 'distance']
	Metric	['euclidean', 'manhattan']

The best parameters for the Decision Tree are as follows: criterion is gini, max_depth is 3, max_features is the square root, min_samples_leaf is 4, and min_samples_split is 2. For Random Forest, the best parameters are n_estimators = 100, min_samples_split = 100, min_samples_leaf = 1, and max_depth = 10. Lastly, for K-Nearest Neighbor, the best parameters are

a Manhattan metric, $n_neighbors$ of 11, and weights set to distance.

3.2 Model Validation

In this study we implement three predictions models. The selected parameters were chosen based on previous studies to ensure optimal model performance [15]. In this study, the FAUS model performance will be compared with the baseline modes. Those baseline models are No Sampling model, Random Under-sampler (RUS) model, and AllKNN under-sampler model. The results of the performance from all models are presented in Table 4. The metrics provide a comprehensive assessment of each model's predictive capability. We present a comparison performance evaluation of the F1-score between sampling techniques in Figure 4.

Table 4. Summary of Evaluation Result

	No Sampling	RUS	AllKNN	FAUS
DT	0.49	0.50	0.48	0.63
RF	0.49	0.50	0.48	0.79
KNN	0.57	0.57	0.57	0.61

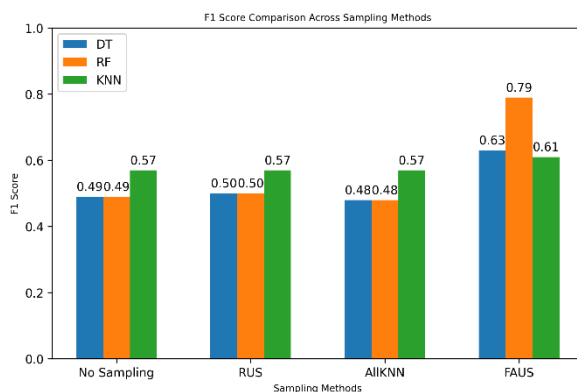


Figure 3. Performance Comparison between Sampling Techniques

The results indicate that FAUS consistently improves model performance RF model achieved the highest F1 score, 0.79. Similarly, the DT with the FAUS model showed a significant improvement in performance, with a score of 0.63, and the KNN with the FAUS model achieved an improving F1 score of 0.61.

The RUS model has similar performance on DT and RF models but is low on the KNN model. This may be due to the randomness in the selection process, which can affect the quality of the selected train set [11]. The No Sampling model achieved a similar F1 Score on DT and RF models, which was not higher than the F1 Score of the KNN model. This model can achieve almost similar results because it doesn't remove any instances [15].

The AllKNN model achieved similar performance on the DT and RF models but high performance on the KNN model. This can occur because of an algorithm based on k selection [13]. The results indicate that each model achieved its highest performance with different sampling techniques. However, the FAUS model still achieved the highest performance evaluation. These results highlight the effectiveness of FAUS in handling

imbalanced data in the Breast Cancer Dataset. This algorithm enhances model performance. The improvement can be attributed to FAUS' ability to optimize instance selection using swarm intelligence, ensuring the informative samples are retained while maintaining the dataset balance.

The FAUS performance is highly dependent on parameter tuning. Improper parameter selection may lead to suboptimal instance selection. Additionally, FAUS has a higher computational complexity compared to traditional under-sampling techniques with an average execution time of approximately 3 seconds per iteration. Future research is needed to explore adaptive parameter tuning and compare FAUS with other metaheuristic-based resampling methods.

4. Conclusions

We have developed the Firefly Algorithm under-sampling to handle the imbalance issue on the Breast Cancer dataset. The results indicate that FAUS consistently improved model performance across various classifiers. For instance, the F1-score of the Random Forest model increased from 0.71 to 0.79, while Decision Tree improved from 0.64 to 0.73, and K-Nearest Neighbors from 0.61 to 0.70, clearly demonstrating the advantage of using FAUS over conventional resampling methods. This enhanced performance can be attributed to the swarm intelligence mechanism in FAUS, which enables the selection of more informative majority class samples. While FAUS showed promising results, further research is needed to benchmark it against other metaheuristics optimization techniques.

References

- [1] N. Tzenios, M. E. Tazanios, and M. Chahine, "The impact of BMI on breast cancer – an updated systematic review and meta-analysis," *Medicine*, vol. 103, no. 5, p. e36831, Feb. 2024, doi: [10.1097/MD.00000000000036831](https://doi.org/10.1097/MD.00000000000036831).
- [2] S. Palmal, N. Arya, S. Saha, and S. Tripathy, "Breast cancer survival prognosis using the graph convolutional network with Choquet fuzzy integral," *Sci Rep*, vol. 13, no. 1, p. 14757, Sep. 2023, doi: [10.1038/s41598-023-40341-z](https://doi.org/10.1038/s41598-023-40341-z).
- [3] N. Adam and R. Wieder, "Predictive Modeling of Long-Term Survivors with Stage IV Breast Cancer Using the SEER-Medicare Dataset," *Cancers*, vol. 16, no. 23, p. 4033, Dec. 2024, doi: [10.3390/cancers16234033](https://doi.org/10.3390/cancers16234033).
- [4] S. Zarean Shahraki *et al.*, "Time-related survival prediction in molecular subtypes of breast cancer using time-to-event deep-learning-based models," *Front. Oncol.*, vol. 13, p. 1147604, Jun. 2023, doi: [10.3389/fonc.2023.1147604](https://doi.org/10.3389/fonc.2023.1147604).
- [5] R. Shafique *et al.*, "Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning," *Cancers*, vol. 15, no. 3, p. 681, Jan. 2023, doi: [10.3390/cancers15030681](https://doi.org/10.3390/cancers15030681).
- [6] G. Aguiar, B. Krawczyk, and A. Cano, "A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework," *Mach Learn*, vol. 113, no. 7, pp. 4165–4243, Jul. 2024, doi: [10.1007/s10994-023-06353-6](https://doi.org/10.1007/s10994-023-06353-6).
- [7] M. Han, A. Li, Z. Gao, D. Mu, and S. Liu, "Hybrid Sampling and Dynamic Weighting-Based Classification Method for Multi-Class Imbalanced Data Stream," *Applied Sciences*, vol. 13, no. 10, p. 5924, May 2023, doi: [10.3390/app13105924](https://doi.org/10.3390/app13105924).

- [8] A. Y. Taha, S. Tiun, A. H. Abd Rahman, and A. Sabah, "Multilabel Over-Sampling And Under-Sampling With Class Alignment For Imbalanced Multilabel Text Classification," *Journal of Information and Communication Technology*, vol. 20, 2021, doi: [10.32890/jict2021.20.3.6](https://doi.org/10.32890/jict2021.20.3.6).
- [9] V. K. Chennuru and S. R. Timmappareddy, "Simulated annealing based undersampling (SAUS): a hybrid multi-objective optimization method to tackle class imbalance," *Appl Intell*, vol. 52, no. 2, pp. 2092–2110, Jan. 2022, doi: [10.1007/s10489-021-02369-4](https://doi.org/10.1007/s10489-021-02369-4).
- [10] I. Chhillar and A. Singh, "An improved soft voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing," *Discov Artif Intell*, vol. 5, no. 1, p. 4, Jan. 2025, doi: [10.1007/s44163-025-00224-w](https://doi.org/10.1007/s44163-025-00224-w).
- [11] D. Namira Purba, G. Khoerunnisa, and I. Kurniawan, "Firefly Algorithm Under-Sampling (FAUS) in Handling Imbalance Data on Drug Target Interaction Prediction," in *2024 12th International Conference on Information and Communication Technology (ICoICT)*, Bandung, Indonesia: IEEE, Aug. 2024, pp.73–78,doi: [10.1109/ICoICT61617.2024.10698356](https://doi.org/10.1109/ICoICT61617.2024.10698356).
- [12] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine*, vol. 103, no. 3, p. e36905,Jan.2024,doi: [10.1097/MD.00000000000036905](https://doi.org/10.1097/MD.00000000000036905).
- [13] M. A. El-Shorbagy and A. M. El-Refaey, "A hybrid genetic–firefly algorithm for engineering design problems," *Journal of Computational Design and Engineering*, vol. 9, no. 2, pp. 706–730,Apr.2022,doi: [10.1093/jcde/qwac013](https://doi.org/10.1093/jcde/qwac013).
- [14] N. Bacanin, K. Venkatachalam, T. Bezdan, M. Zivkovic, and M. Abouhawwash, "A novel firefly algorithm approach for efficient feature selection with COVID-19 dataset," *Microprocessors and Microsystems*, vol. 98, p. 104778, Apr. 2023, doi: [10.1016/j.micpro.2023.104778](https://doi.org/10.1016/j.micpro.2023.104778).
- [15] D. N. Purba, Fhira Nhita, and Isman Kurniawan, "Implementation of Ensemble Method in Schizophrenia Identification Based on Microarray Data," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 6, no. 1, pp. 64–69, Feb. 2022, doi: [10.29207/resti.v6i1.3788](https://doi.org/10.29207/resti.v6i1.3788).
- [16] A. Amkor and N. El Barbri, "Classification of potatoes according to their cultivated field by SVM and KNN approaches using an electronic nose," *Bulletin EEI*, vol. 12, no. 3, pp. 1471–1477, Jun. 2023, doi: [10.11591/eei.v12i3.5116](https://doi.org/10.11591/eei.v12i3.5116).
- [17] I. Mettildha Mary and K. Karuppasamy, "Adaptive Kernel Firefly Algorithm Based Feature Selection and Q-Learner Machine Learning Models in Cloud," *Computer Systems Science and Engineering*, vol. 46, no. 3, pp. 2667–2685, 2023, doi: [10.32604/csse.2023.031114](https://doi.org/10.32604/csse.2023.031114).
- [18] X.-S. Yang, "Firefly Algorithm, Stochastic Test Functions and Design Optimisation," Mar. 06, 2010, *arXiv:1003.1409*. Accessed: Jun. 06, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.1003.1409>.
- [19] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: [10.1109/ACCESS.2021.3102399](https://doi.org/10.1109/ACCESS.2021.3102399).
- [20] V. Tomar, M. Bansal, and P. Singh, "Metaheuristic Algorithms for Optimization: A Brief Review," in *RAiSE-2023, MDPI*, Mar.2024,p.238,doi: [10.3390/engproc2023059238](https://doi.org/10.3390/engproc2023059238).
- [21] P. H. Putra, A. Azanuddin, B. Purba, and Y. A. Dalimunthe, "Random forest and decision tree algorithms for car price prediction," *JUMPA*, vol. 4, no. 1, pp. 81–89, Sep. 2023, doi: [10.54076/jumpa.v3i2.305](https://doi.org/10.54076/jumpa.v3i2.305).
- [22] M. Fayaz, A. Abadi, and S. Khodakarim, "Random-Splitting Random Forest with Multiple Mixed-Data Covariates," *jbe*, Oct. 2023, doi: [10.18502/jbe.v9i1.13974](https://doi.org/10.18502/jbe.v9i1.13974).
- [23] Al Danny Rian Wibisono, Syahrul Hidayat, Humam Maulana Tsubasanofa Ramadhan, and Eva Yulia Puspaningrum, "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," *ijodas*, vol. 4, no. 2, pp. 90–100, Jul. 2023, doi: [10.56705/ijodas.v4i2.70](https://doi.org/10.56705/ijodas.v4i2.70).
- [24] K. Eldora, E. Fernando, and W. Winanti, "Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction: A Machine Learning Approach," *J. Inf. Syst. Informatics*, vol. 6, no. 1, pp. 313–338, Mar. 2024, doi: [10.51519/journalisi.v6i1.664](https://doi.org/10.51519/journalisi.v6i1.664).
- [25] K. M. Almustafa, "Prediction of chronic kidney disease using different classification algorithms," *Informatics in Medicine Unlocked*, vol. 24, p. 100631, 2021, doi: [10.1016/j.imu.2021.100631](https://doi.org/10.1016/j.imu.2021.100631).