Optimization of Heart Failure Classification on Imbalanced Data Using a Supervised Learning Approach Based on Logistic Regression, Random Forest, and K-Nearest Neighbor

Feri Agustina[1], Candra Irawan[2], Lalang Erawan[3], Suprayogi[4], Deddy Award Widya Laksana[5], Cahaya Jatmoko[6], Daurat Sinaga[7], Heru Lestiawan[8], Mohamed Doheir[9]

[1,2,3,4,5,6,7,8] Informatics Engineering, Computer Science, Dian Nuswantoro University, Indonesia
[9] Department of Technology Management, Faculty of Technology Management and Technopreneurship, Universiti Teknikal, Malaysia

[1] feri.agustina@dsn.dinus.ac.id, [2] candra.irawan@dsn.dinus.ac.id, [3] lalang.erawan@dsn.dinus.ac.id, [4] suprayogismg@gmail.com, [5] deddyawardwidyalaksana@gmail.com, [6] cahayajatmoko@dsn.ac.id, [7] dauratsinaga@dsn.dinus.ac.id, [8] Heru.lestiawan@dsn.dinus.ac.id

Abstract

Heart failure remains one of the leading causes of mortality worldwide, posing significant challenges for early diagnosis and patient management. One of the major obstacles in developing predictive models for heart failure is the class imbalance problem, where the number of surviving patients far exceeds those who experience death events. This imbalance often leads machine learning algorithms to bias toward the majority class, reducing sensitivity to critical minority cases. To address this issue, this study applies the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and improve model performance. Three supervised learning algorithms, namely Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbor (KNN), were implemented and compared on the UCI Heart Failure Clinical Records dataset containing 299 patient samples with 13 clinical attributes. Experimental results show that the Random Forest model achieved the highest performance with 90% accuracy, precision, recall, and F1-score, outperforming both LR and KNN. The findings demonstrate that combining data balancing with ensemble learning effectively enhances prediction accuracy and sensitivity toward minority classes. The main contribution of this research lies in optimizing supervised models for medical data with skewed class distributions, providing a more reliable and interpretable approach for early heart failure detection. Future research may extend this work by integrating advanced ensemble or hybrid deep learning models and expanding the dataset for multi-institutional validation.

Keywords: Class Imbalance, Heart Failure, K-Nearest Neighbor, Logistic Regression, Random Forest, SMOTE.

1. Introduction

Heart failure is one of the leading causes of death and a major burden on public health systems worldwide (Bhatt et al., 2023; Chandrasekhar & Peddakrishna, 2023). According to the World Heart Federation report (2023), cardiovascular diseases, including heart failure, cause approximately 20 million deaths annually, accounting for nearly one in three deaths globally (World Health Organization, 2024). It is estimated that more than 64 million people live with heart failure worldwide. In the context of applying data mining and machine learning for early detection of heart failure, a serious challenge arises in the form of class imbalance, where the number of patients with heart failure (minority class) is significantly lower than that of healthy patients (majority class) (Jaddoa, 2023). This imbalance causes supervised learning algorithms to be biased toward the majority class, thereby reducing sensitivity to the minority class and increasing the risk of misclassification (Amirruddin et al., 2022; Muzakki et al., 2023). Therefore, research on optimizing heart failure classification with class imbalance handling is crucial to improve early detection accuracy and help reduce the global mortality rate associated with heart failure.

As a solution to the class imbalance problem in heart failure data, an optimization-based supervised learning approach can be used to enhance classification performance, particularly in detecting high-risk minority cases (Cahyo et al., 2023; Farhan et al., 2023; Kamila et al., 2023). his strategy can be implemented through a combination of data preprocessing techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) to balance class distribution, and the selection of algorithms that are adaptive to medical data, such as Logistic Regression (Šinkovec et al., 2021), Random Forest (Rasyidi et al., 2021), and K-Nearest Neighbor (Hasanah et al., 2024). Logistic Regression offers high interpretability in clinical contexts, Random

Forest effectively handles non-linear relationships and class imbalance through the bagging mechanism, while KNN is effective in recognizing local patterns between patient instances. Moreover, the use of cross-validation and evaluation metrics such as precision, recall, and F1-score ensures the model's stability on imbalanced datasets (Basha et al., 2022). With this approach, the heart failure classification system can be optimized to deliver more accurate early detection and potentially reduce global mortality due to heart failure.

Research by (Sabouri et al., 2023) utilized data from 737 heart failure patients from the RASHF registry to predict in-hospital mortality, six-month mortality, and 30-day and 90-day readmission. The methodology included Z-score normalization, three feature selection techniques (Boruta, RFE, MRMR), eight ML algorithms, hyperparameter optimization with cross-validation, and 1,000 bootstrap iterations on the hold-out set. The best performance for in-hospital mortality was achieved by the LR model with RFE (AUC 0.91, accuracy 0.84, sensitivity 0.83, specificity 0.84). However, for 3-month readmission and 6-month mortality, the performance was relatively low (e.g., AUC ~0.60 for 3-month readmission). The main limitations were the single-center dataset and reduced performance for medium/long-term outcomes, indicating challenges in generalization and handling minority-class outcomes.

Study by (Moreno-Sánchez, 2023) proposed a data optimization pipeline that integrates feature selection, algorithm selection, and explainable AI (XAI) analysis for predicting survival in 299 heart failure patients. Two approaches were built: one for survival analysis (Gradient Boosting survival model with c-index 0.714) and one for classification (Random Forest with balanced accuracy ~0.74 ± 0.03). The most important identified features included serum_creatinine, ejection_fraction, and gender. Limitations included a small sample size and single cohort, as well as the retrospective nature of the data, which limits generalization to broader populations.

Study by (Li et al., 2023) developed a deep learning system based on CNN with multi-head self-attention to predict four categories of mortality in heart failure patients (death within 30 days, 180 days, 365 days, and after 365 days) using the public MIMIC-III database (10,311 patients). To address class imbalance, they used focal loss. The results showed that the system effectively predicted all four mortality categories and applied Deep SHAP for feature interpretability. However, despite the large dataset, deep learning models tend to be "black boxes" and still require external validation in different populations, as well as significant computational resources and data demands.

Based on previous studies, there remains a significant research gap in developing a heart failure classification model that achieves a balance between high predictive accuracy and clinical interpretability on imbalanced data. Therefore, the novelty of this study lies in the development of an optimized heart failure classification approach based on supervised learning that combines Logistic Regression, Random Forest, and K-Nearest Neighbor algorithms integrated with the SMOTE data balancing technique to enhance model performance and sensitivity toward the minority class while maintaining interpretability within a medical context.

2. Methods

As seen in Figure 1, the process begins with the raw dataset, which undergoes a pre-processing phase that includes data cleaning to remove missing values, duplicates, and inconsistencies, ensuring that the data is ready for model training. Afterward, the dataset is split into two subsets, with 80% used for training and 20% for testing. To address the problem of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is applied to the training data to synthetically generate minority-class samples, resulting in a more balanced class distribution.

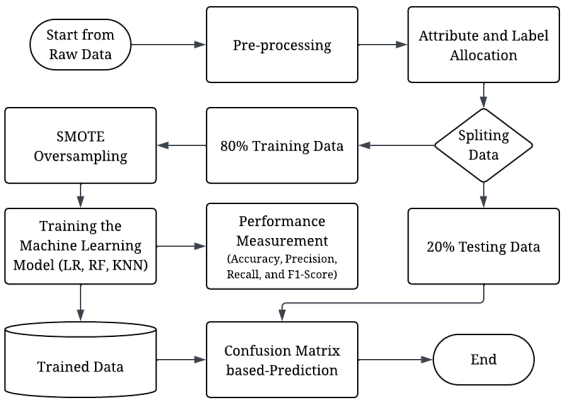


Figure 1. Flow of Proposed Methods

Next, three supervised machine learning algorithms namely LR, RF, and KNN are trained on the balanced training data to develop predictive models. The trained models are then evaluated using the testing dataset, where performance metrics such as accuracy, precision, recall, and F1-score are measured to assess classification quality. Additionally, a confusion matrix is generated to analyze the model's ability to correctly distinguish between positive and negative cases. The process concludes with the creation of an optimized trained model capable of performing accurate and reliable early detection of heart failure, even when dealing with imbalanced datasets.

2.1 Data Collections (Raw Data)

The dataset used in this study was obtained from the UCI Machine Learning Repository, available at

https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records. This dataset contains the medical records of 299 patients who experienced heart failure during a follow-up period. Each patient profile includes 13 clinical features that represent demographic, laboratory, and clinical measurements associated with cardiovascular conditions. These attributes serve as predictors, while the target variable (DEATH_EVENT) indicates whether the patient died during the follow-up period.

The dataset is structured without missing values and consists of both categorical (boolean/binary) and continuous numerical variables. The features capture vital health indicators such as age, blood chemistry, cardiac ejection fraction, comorbidities (e.g., anaemia, diabetes, hypertension), and lifestyle factors like smoking. The overall goal is to classify patients into survival or death outcomes based on these medical parameters.

Table 1. Raw data overview

Variable Name	*Type*	*Description*	*Units*	*Missing Value*
age	Integer	Age of the patient	years	No
anaemia	Binary	Decrease of red blood cells or hemoglobin	Boolean	No
creatinine_phosphokinase	Integer	Level of the CPK enzyme in the blood	mcg/L	No
diabetes	Binary	Whether the patient has diabetes	Boolean	No
ejection_fraction	Integer	Percentage of blood leaving the heart at each contraction	%	No
high_blood_pressure	Binary	Whether the patient has hypertension	Boolean	No
platelets	Continuous	Platelet count in the blood	kiloplatelets/mL	No
serum_creatinine	Continuous	Level of serum creatinine in the blood	mg/dL	No
serum_sodium	Integer	Level of serum sodium in the blood	mEq/L	No
sex	Binary	Gender (male or female)	Binary	No
smoking	Binary	Whether the patient smokes	Boolean	No
time	Integer	Follow-up period	days	No
death_event (target)	Binary	Indicates if the patient died during follow-up	Boolean	No

As seen in Table 1, the dataset provides a comprehensive collection of clinical, demographic, and laboratory variables that are highly relevant for predicting heart failure outcomes. The combination of both continuous and categorical features enables the construction of diverse supervised learning models that can capture non-linear relationships between medical indicators and patient survival. Since all attributes are well-defined and contain no missing values, the dataset is ideal for developing machine learning models without requiring extensive data imputation or correction. Overall, the dataset's structure and feature diversity make it suitable for implementing classification optimization through techniques such as SMOTE and multiple supervised algorithms to improve prediction accuracy and model generalization.

2.2 Pre-processing

In this stage, the pre-processing phase was relatively straightforward because the dataset did not contain any missing values or duplicate entries (Palanivinayagam & Damaševičius, 2023). The main focus of this step was to ensure that all variables had the correct data types corresponding to their roles before moving on to the model training phase. Numerical attributes such as age, ejection_fraction, serum_creatinine, and platelets were converted into numeric formats suitable for analysis, while categorical features such as sex, anaemia, diabetes, and smoking were encoded as binary values (0 and 1). The dataset was then separated into attributes (independent features) and the label (target variable), where the target column DEATH_EVENT represented patient survival status. This pre-processing ensured data consistency and compatibility for the subsequent SMOTE oversampling and supervised learning model training stages.

2.3 Synthetic Minority Oversampling Technique

To address the issue of class imbalance in the heart failure dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data before the model training phase (Adi Pratama & Oktora, 2023). The original dataset exhibited an unequal distribution between patients who survived (majority class) and those who experienced death events (minority class), which could bias the learning process of supervised algorithms and reduce sensitivity toward the minority class. SMOTE helps mitigate this problem by generating synthetic samples of the minority class rather than simply duplicating existing ones. It does so by interpolating between a randomly selected minority instance and its nearest neighbors within the feature space, thereby creating new, realistic data

points that enrich the minority class representation (Hasanah et al., 2024; Muzakki et al., 2023).

The solution by applying SMOTE, the class distribution in the training dataset becomes more balanced, allowing algorithms such as Logistic Regression, Random Forest, and K-Nearest Neighbor to learn the classification boundaries more effectively. This balance enhances the model's ability to detect patients at higher risk while minimizing bias toward the majority group. The improved distribution also contributes to better recall and F1-score values, ensuring that the predictive model is not only accurate but also sensitive to critical cases of heart failure.

Figure 2. Class distribution of the death event variable before and after applying SMOTE

The effect of applying the SMOTE technique on the dataset distribution is illustrated in Figure 2. The left chart shows the class imbalance before oversampling, where the number of patients who survived (class 0) was significantly higher than those who experienced a death event (class 1). After applying SMOTE, as shown in the right chart, both classes became approximately balanced. This equalized distribution enables the supervised learning models to train more effectively without bias toward the majority class, leading to improved overall classification performance and better detection of minority-class patients.

2.4 Machine Learning Models

This study employed three supervised learning algorithms: LR, RF, and KNN for classifying heart failure outcomes. Each algorithm was trained on the balanced dataset obtained after applying SMOTE and optimized through cross-validation. These algorithms were chosen to represent linear, ensemble, and instance-based learning approaches, respectively, ensuring a comprehensive performance comparison.

2.4.1 Logistic Regression (LR)

LR is a statistical model used for binary classification that estimates the probability of a categorical outcome based on a set of predictor variables (Šinkovec et al., 2021). The model applies the sigmoid activation function to map outputs between 0 and 1, representing the probability of a patient experiencing a death event. Its simplicity and interpretability make it suitable for clinical analysis, allowing researchers to understand how each clinical variable contributes to the prediction outcome.

In this study, Logistic Regression was implemented using the liblinear solver with L2 regularization to prevent overfitting and ensure convergence. The parameter configuration used in this research is presented in Table 2.

Table 2. Parameter of Logistic Regression

Parameter	*Value*	*Description*
solver	liblinear	Optimization algorithm suitable for small and binary datasets
penalty	l2	Ridge regularization to reduce overfitting
max_iter	1000	Maximum number of iterations for convergence
random_state	42	Ensures reproducibility of results

2.4.2 Random Forest (RF)

RF is an ensemble-based supervised learning algorithm that constructs multiple decision trees during training and combines their outputs through majority voting to achieve higher prediction accuracy and robustness (Daviran et al., 2023; Teodorescu & Obreja Brașoveanu, 2025). It is effective in handling complex, non-linear relationships and reducing overfitting through random sampling of both features and data subsets (bagging) (Albert et al., 2022).

For this study, Random Forest was applied to the balanced dataset generated after SMOTE, with the key hyperparameters tuned to optimize predictive performance. The complete list of parameters used for the Random Forest model is summarized in Table 3.

Table 3. Parameter of Random Forest

Parameter	*Value*	*Description*
n_estimators	100	Number of trees in the forest
criterion	gini	Splitting criterion to measure node purity
max_depth	None	Allows full tree growth for capturing complex patterns
min_samples_split	2	Minimum number of samples required to split a node
random_state	42	Ensures reproducibility and consistent results

2.4.3 K-Nearest Neighbor (KNN)

KNN is a non-parametric, distance-based learning algorithm that classifies new data points according to the majority class among their k closest neighbors in the feature space (Ab Wahab et al., 2021; Irawan et al., 2021). Unlike parametric models, KNN does not make assumptions about data distribution, which allows it to model non-linear decision boundaries effectively. Its performance heavily

depends on the choice of k and the distance metric used.

In this research, KNN was applied to evaluate how neighborhood-based classification performs on balanced medical data after SMOTE. The parameters selected for this study were determined empirically to balance between bias and variance, as shown in Table 4.

Table 4. Parameter of KNN

Parameter	*Value*	*Description*
n_neighbors	5	Number of nearest neighbors considered for classification
metric	euclidean	Distance metric used to measure similarity between data points
weights	uniform	Assigns equal weight to all neighbors
algorithm	auto	Automatically selects the most efficient computation method

2.5 Matrics Evaluation

To assess the performance of the supervised learning models, four key evaluation metrics were employed: Accuracy, Precision, Recall, and F1-Score (Chicco et al., 2021; Fan, 2025). These metrics were chosen because they provide a comprehensive assessment of model performance, especially when working with imbalanced datasets such as the heart failure data used in this study. While accuracy measures the overall proportion of correct predictions, it may be misleading when class distribution is uneven (Al-Ghiffary et al., 2024). Therefore, precision, recall, and F1-score were also considered to better capture the model's sensitivity and reliability in identifying the minority class (death events).

$$Accuracy = \frac{TP+TN}{All\ Data} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{recision+Recall} \quad (4)$$

As shown in the formulas above, Accuracy quantifies the overall correctness of model predictions, while Precision focuses on how many predicted positives are actually correct. Recall measures the model's ability to detect all actual positive cases, which is crucial in medical diagnosis tasks where missing a high-risk patient can be critical. Finally, the F1-Score serves as a harmonic mean of precision and recall, providing a balanced evaluation of both false positives and false negatives.

3. Result and Discussions

This section presents the experimental results and analysis obtained from the implementation of the proposed heart failure classification models. All experiments were conducted using the Python programming language, which provides an extensive ecosystem of libraries suitable for data analysis and machine learning. The implementation utilized Jupyter Notebook as the development environment, supported by libraries such as Pandas and NumPy for data manipulation, Scikit-learn (sklearn) for model training, SMOTE balancing, and evaluation, as well as Matplotlib and Seaborn for visualization of results.

3.1 Training Section

After completing data preprocessing and class balancing using SMOTE, each machine learning algorithm was trained using the processed dataset. The training phase aimed to evaluate and compare the predictive performance of the three supervised learning models. Model evaluation was conducted based on four performance metrics: Accuracy, Precision, Recall, and F1-Score, as previously described in Section 2.5. These metrics provide a balanced assessment of model performance, particularly in detecting minority-class cases (patients who experienced death events).

The results of the training process for each model are summarized in Table 5, which displays the comparative performance across all evaluation metrics.

Table 5. Metrics Performance

Model	*ACC*	*Precision*	*Recall*	*F1-Score*
LR	0.80	0.80	0.80	0.80
RF	0.90	0.90	0.90	0.90
KNN	0.80	0.80	0.80	0.80

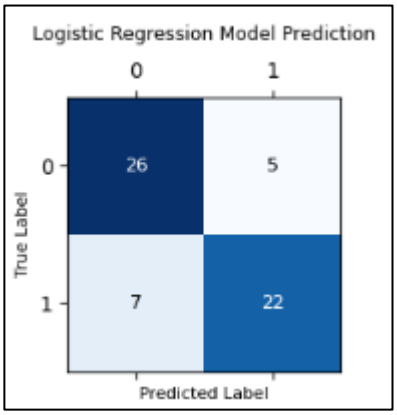
As shown in Table 5, the Random Forest model outperformed the other algorithms, achieving the highest accuracy, precision, recall, and F1-score of 0.90 across all metrics. This indicates that Random Forest was able to effectively capture complex patterns and provide a balanced classification between survival and death events, even under conditions of prior class imbalance. Meanwhile, Logistic Regression and KNN achieved moderate but consistent performance, demonstrating stable predictive ability on the balanced dataset.

Overall, the training results confirm that ensemble-based learning provides superior generalization for heart failure prediction tasks. Following the training stage, the next phase involves testing the trained models on unseen data to evaluate their predictive capability and robustness.

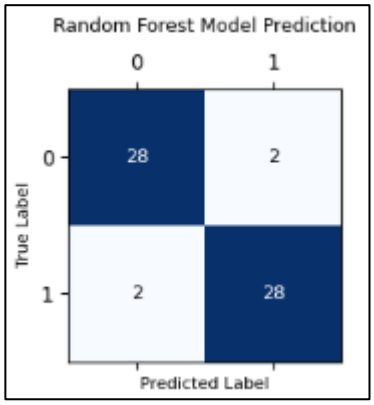
3.2 Model Predictions (Testing Section)

In this section, the performance of the trained models was further evaluated on the testing dataset to analyze their prediction capability and classification behavior. The confusion matrices for each model are illustrated in Figure 3, where Figure 3(a) represents the Logistic Regression model, Figure 3(b) the Random Forest model, and Figure 3(c) the K-Nearest

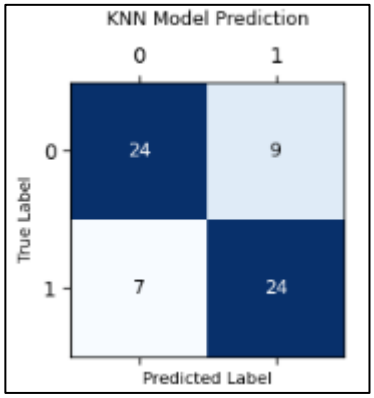
Neighbor model. These matrices provide a detailed view of the correct and incorrect classifications made by each algorithm on unseen data.



(a)



(b)



(c)

Figure 3. Model Prediction Based on Confusion Matrix (a) Logistic Regression Model Prediction, (b) Random Forest Model Prediction, (c) KNN Model Prediction

As shown in Figure 3(a), the Logistic Regression (LR) model correctly classified most samples but still produced a total of 12 misclassifications, consisting of 5 false positives (class 0 predicted as 1) and 7 false negatives (class 1 predicted as 0). This result indicates that although LR performs consistently, it shows limited sensitivity in detecting minority cases (death events), which may be due to its linear decision boundary and inability to capture complex feature interactions.

In Figure 3(b), the Random Forest (RF) model achieved the best performance, producing only 4 misclassifications in total, comprising 2 false positives and 2 false negatives. The RF model's ensemble structure allows it to effectively capture non-linear patterns and interactions among clinical features, resulting in higher accuracy, balanced precision–recall, and stronger robustness compared to single-model classifiers.

Meanwhile, Figure 3(c) shows that the K-Nearest Neighbor (KNN) model achieved moderate predictive performance, with 16 total misclassifications, including 9 false positives and 7 false negatives. The relatively lower performance can be attributed to KNN's sensitivity to feature scaling and overlapping neighborhood regions, which may cause confusion between similar patient samples, particularly in datasets with mixed numerical and binary attributes.

Overall, these results indicate that the Random Forest model achieved the most reliable classification on the testing set, followed by Logistic Regression, while KNN exhibited the weakest performance due to its higher rate of misclassification.

4. Conclusions

This research addressed the challenge of class imbalance in heart failure classification, a common issue that often causes predictive models to perform poorly in detecting minority cases such as death events. By applying the Synthetic Minority Oversampling Technique (SMOTE), the imbalance between the survival and death classes was effectively mitigated, enabling a more balanced learning process. Three supervised learning algorithms—Logistic Regression, Random Forest, and K-Nearest Neighbor—were implemented to evaluate model performance on the balanced dataset. The results demonstrated that the Random Forest model achieved the highest overall performance with 90% accuracy, precision, recall, and F1-score, outperforming Logistic Regression and KNN. This indicates that ensemble-based learning methods can handle complex clinical data patterns more effectively, providing robust and interpretable outcomes for heart failure prediction.

The main contribution of this study lies in the integration of class balancing techniques and model optimization to enhance predictive accuracy on medical datasets with uneven distributions. The approach demonstrated how combining traditional supervised learning algorithms with oversampling techniques can improve model fairness and clinical reliability. For future research, it is recommended to explore advanced ensemble or hybrid deep learning models such as XGBoost, LightGBM, or neural network architectures, along with feature selection and interpretability frameworks like SHAP or LIME. Additionally, expanding the dataset with multi-

hospital or real-time electronic health record data could further improve model generalization and applicability in clinical decision-support systems.

References

Ab Wahab, M. N., Nazir, A., Ren, A. T. Z., Noor, M. H. M., Akbar, M. F., & Mohamed, A. S. A. (2021). Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi. *IEEE Access*, *9*, 134065–134080. https://doi.org/10.1109/ACCESS.2021.3113337

Adi Pratama, F. R., & Oktora, S. I. (2023). Synthetic Minority Over-sampling Technique (SMOTE) for handling imbalanced data in poverty classification. *Statistical Journal of the IAOS*, *39*(1), 233–239. https://doi.org/10.3233/SJI-220080

Albert, A. J., Murugan, R., & Sripriya, T. (2022). Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. *Research on Biomedical Engineering*, *39*(1), 99–113. https://doi.org/10.1007/s42600-022-00253-9

Al-Ghiffary, M. M. I., Cahyo, N. R. D., Rachmawanto, E. H., Irawan, C., & Hendriyanto, N. (2024). Adaptive deep learning based on FaceNet convolutional neural network for facial expression recognition. *Journal of Soft Computing*, *05*(03), 271–280. https://doi.org/https://doi.org/10.52465/joscex.v5i3.450

Amirruddin, A. D., Muharam, F. M., Ismail, M. H., Tan, N. P., & Ismail, M. F. (2022). Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)-Adaptive Boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (Elaeis guineensis) using spectroradiometers and unmanned aerial vehicles. *Computers and Electronics in Agriculture*, *193*, 106646. https://doi.org/10.1016/j.compag.2021.106646

Basha, S. J., Madala, S. R., Vivek, K., Kumar, E. S., & Ammannamma, T. (2022). A Review on Imbalanced Data Classification Techniques. *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, 1–6. https://doi.org/10.1109/ICACTA54488.2022.9753392

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2). https://doi.org/10.3390/a16020088

Cahyo, N. R. D., Sari, C. A., Rachmawanto, E. H., Jatmoko, C., Al-Jawry, R. R. A., & Alkhafaji, M. A. (2023). A Comparison of Multi Class Support Vector Machine vs Deep Convolutional Neural Network for Brain Tumor Classification. *2023 International Seminar on Application for Technology of Information and Communication (ISemantic)*, *12*(2), 358–363. https://doi.org/10.1109/iSemantic59612.2023.10295336

Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4). https://doi.org/10.3390/pr11041210

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*(1), 13. https://doi.org/10.1186/s13040-021-00244-z

Daviran, M., Shamekhi, M., Ghezelbash, R., & Maghsoudi, A. (2023). Landslide susceptibility prediction using artificial neural networks, SVMs and random forest: hyperparameters tuning by genetic optimization algorithm. *International Journal of Environmental Science and Technology*, *20*(1), 259–276. https://doi.org/10.1007/s13762-022-04491-3

Fan, C.-L. (2025). Evaluation Model for Crack Detection with Deep Learning: Improved Confusion Matrix Based on Linear Features. *Journal of Construction Engineering and Management*, *151*(3). https://doi.org/10.1061/JCEMD4.COENG-14976

Farhan, F., Sari, C. A., Rachmawanto, E. H., & Cahyo, N. R. D. (2023). Mangrove Tree Species Classification Based on Leaf, Stem, and Seed Characteristics Using Convolutional Neural Networks with K-Folds Cross Validation Optimalization. *Advance Sustainable Science Engineering and Technology*, *5*(3), 02303011. https://doi.org/10.26877/asset.v5i3.17188

Hasanah, U., Soleh, A. M., & Sadik, K. (2024). Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models. *Jurnal Matematika, Statistika Dan Komputasi*, *21*(1), 88–102. https://doi.org/10.20956/j.v21i1.35552

Irawan, C., Winarno, A., Kusumodestoni, H., Sucipto, A., Tamrin, T., & Doheir, M. (2021). A Combination of Statistical Extraction and Texture Features Based on KNN for Batik Classification. *2021 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 113–117. https://doi.org/10.1109/iSemantic52711.2021.9573214

Jaddoa, A. S. (2023). *Heart disease prediction system using (SMOTE technique) balanced dataset and decision tree classifier*. 050006. https://doi.org/10.1063/5.0161558

Kamila, I. P., Sari, C. A., Rachmawanto, E. H., & Cahyo, N. R. D. (2023). A Good Evaluation Based on Confusion Matrix for Lung Diseases Classification using Convolutional Neural Networks. *Advance Sustainable Science, Engineering and Technology*, *6*(1), 0240102. https://doi.org/10.26877/asset.v6i1.17330

Li, D., Fu, J., Zhao, J., Qin, J., & Zhang, L. (2023). A deep learning system for heart failure mortality prediction. *PLOS ONE*, *18*(2), e0276835. https://doi.org/10.1371/journal.pone.0276835

Moreno-Sánchez, P. A. (2023). Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Frontiers in Cardiovascular Medicine*, *10*. https://doi.org/10.3389/fcvm.2023.1219586

Muzakki, M. F., Prayogo, R. D., & Rizky A, M. A. (2023). Handling Imbalanced Data for Acute Coronary Syndrome Classification Based on Ensemble and K-Means SMOTE Method. *JOIV : International Journal on Informatics Visualization*, *7*(3–2), 1989. https://doi.org/10.30630/joiv.7.3-2.1429

Palanivinayagam, A., & Damaševičius, R. (2023). Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods. *Information*, *14*(2), 92. https://doi.org/10.3390/info14020092

Rasyidi, M. A., Bariyah, T., Riskajaya, Y. I., & Septyani, A. D. (2021). Classification of handwritten javanese script using random forest algorithm. *Bulletin of Electrical Engineering and Informatics*, *10*(3), 1308–1315. https://doi.org/10.11591/eei.v10i3.3036

Sabouri, M., Rajabi, A. B., Hajianfar, G., Gharibi, O., Mohebi, M., Avval, A. H., Naderi, N., & Shiri, I. (2023). Machine learning based readmission and mortality prediction in heart failure patients. *Scientific Reports*, *13*(1), 18671. https://doi.org/10.1038/s41598-023-45925-3

Šinkovec, H., Heinze, G., Blagus, R., & Geroldinger, A. (2021). To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Medical Research Methodology*, *21*(1), 199. https://doi.org/10.1186/s12874-021-01374-y

Teodorescu, V., & Obreja Brașoveanu, L. (2025). Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost. *Computation*, *13*(5), 127. https://doi.org/10.3390/computation13050127

World Health Organization. (2024, September 29). *World Heart Day: Cardiovascular diseases claim 3.9 million lives in the WHO South-East Asia Region every year*. https://www.who.int/southeastasia/news/detail/29-09-2024-world-heart-day