Enhancing Single Nucleotide Polymorphisms Detection from Imbalanced Data: A Study of Resampling Techniques in Machine Learning Algorithms

[1]Rossy Nurhasanah, [2]Dedy Arisandi, [3]Fanindia Purnamasari,
[4]Hayatunnufus, [5]Daisy Sere Damara Simangunsong, [6]Aflah Mutsanni Pulungan
[1,2,3,5,6]Department of Information Technology, Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Sumatera Utara, Indonesia
[4]Department of Computer Science, Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Sumatera Utara, Indonesia
Email: [1]rossynurhasanah@usu.ac.id, [2]dedyarisandi@usu.ac.id, [3]fanindia@usu.ac.id,
[4]hayatunnufus@usu.ac.id, [5]daisyseredams@gmail.com, [6]aflahmutsannipulungan@gmail.com

Article Info

Article history:
Received Oct 6th, 2024
Revised Feb 28th, 2025
Accepted Mar 18th, 2025

Keyword:
Artificial Neural Network
Borderline-SMOTE
Imbalanced Classification
Random Forest
SNP Identification

ABSTRACT

Identifying the actual Single Nucleotide Polymorphisms (SNPs) by sourcing Next Generation Sequencing (NGS) data emerges an imbalanced problem due to the inherent high error rate of NGS technology. The imbalance problem has been found to have a negative impact on machine learning algorithms because it produces biased models and poor performance, particularly in detecting actual SNP that belong to the underrepresented class in question. This study evaluates the effectiveness of several resampling techniques, including Borderline-SMOTE, Random Undersampling, and Tomek-Link, in enhancing the performance of machine learning algorithms, specifically Random Forest (RF) and Artificial Neural Networks (ANN). Furthermore, we compare these techniques to determine the most effective approach. Our results indicate that Borderline-SMOTE improves the F-Measure of RF from 69.72 to 91.52 (a 31.2% increase) and ANN from 79.75 to 91.32 (a 14.5% increase) and outperforms other resampling methods. These findings highlight the crucial role of resampling techniques and the careful selection of algorithms in improving classification accuracy for imbalanced datasets.

Corresponding Author:
Rossy Nurhasanah
Departement of Information Technologi
Universitas Sumatera Utara
Jl Dr. T. Mansur No.9, Kota Medan, Sumatera Utara 20155, Indonesia
Email: rossynurhasanah@usu.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i1.32942

1. INTRODUCTION

Single Nucleotide Polymorphisms (SNPs) can lead to genetic differences, occurring when a single nucleotide A, C, T, or G in the Deoxyribonucleic Acid (DNA) sequence differs among individuals [1]. SNP analysis is crucial in genetics research, with implications covering medicine, agriculture, and forensic science [2]. In humans, SNPs act as genetic markers, offering insights into various diseases and traits. They aid in identifying disease-associated genes and accelerate personalized drug development [3]. In agriculture, SNP mining in plant genomes improves breeding efficiency and serves as diagnostic markers for precise pathogen identification and disease management [4][5]. In forensics, downstream analysis of SNPs, specifically Microhaplotypes (MHs), serves as potential forensic markers for tasks like individual identification, kinship analysis, and lineage prediction [6]. Overall, the demand for valid SNP data is increasing rapidly due to their significant roles across multiple fields.

Using machine learning to identify valid SNPs shows promise but faces challenges due to the high error rates in Next Generation Sequencing (NGS) data as the source of SNP mining. Those errors can arise from sequencing process, alignment issues, or inadequate data coverage [7]. During multiple sequence alignment with the reference genome, most variations found are errors rather than valid SNPs that leads to an imbalanced data problem. Using imbalanced data to build a classifier leads to overfitting to the majority class during training and impacting downstream analyses [8][9]. Furthermore, limited exposure to minority class data may lead the model to overlook them and treat them as noise [10]. Consequently, the model may struggle to accurately classify minority class instances and frequently misclassifying them as the dominant class [11]. Thus, a robust method is needed to accurately distinguish valid SNPs from errors.

Various methods can address imbalanced data in classification tasks, with algorithms such as Random Forest (RF) and Artificial Neural Networks (ANN) renowned for achieving strong classification performance in such challenges [12][13]. RF is considered effective in handling imbalanced datasets, whether they are binary or multiclass, because of its ensemble approach. In a study by [14], patient data was utilized to train RF models for predicting the risk of chronic illnesses based on medical records. The RF method outperforms Support Vector Machine (SVM), bagging, and boosting in predicting disease risk from highly imbalanced data, with an average Area Under the Curve (AUC) of 88.79%. In another study involving individuals without diagnosed coronary artery disease, RF with Synthetic Minority Over-sampling Technique (SMOTE) sampling achieved the highest AUC of 0.97, surpassing all other models [15]. Besides RF, ANN also renown to has a good performance in imbalanced dataset. In a study conducted by [16], ANN were used to predict fetal outcomes within the context of Systemic Lupus Erythematosus (SLE). The rarity of pregnancy among SLE patients, due to its low global prevalence, limits data-driven model predictions and creates an imbalance issue in machine learning. A well-trained ANN has a high sensitivity of 19/21 (90.8%) for identifying patients with fetal loss outcomes in SLE pregnancies. Another study focuses on addressing imbalanced data challenges using ANN in the context of network intrusion detection. The study demonstrates that ANN exhibit improved classification performance when applied to imbalanced data for network intrusion detection, especially after employing resampling techniques to rebalance the dataset, the ANN achieve better accuracy in classifying intrusion events [17].

A part from selecting a suitable machine learning approach, another important strategy to solve imbalanced problem involves utilizing resampling techniques, either through oversampling or undersampling, with the aim of achieving a more balanced dataset. Oversampling methods typically involve duplicating existing samples, creating synthetic samples using techniques like SMOTE, or generating new samples using generative models [18]. Conversely, undersampling techniques entail either removing certain samples from the majority class or merging similar samples. However, both methods have their potential drawbacks such as oversampling may lead to overfitting and increased data complexity while undersampling could lead to the elimination of pertinent data associated with the majority class [19]. The choice between these two methods should consider the data characteristics and the specific goals of the classification analysis. In certain cases, a combination of both approaches and other techniques may be employed to achieve a better balance between classes [20], [21].

Based on several prior research, SMOTE emerges as among the widely utilized oversampling procedure. SMOTE synthesizes augmented samples for the underrepresented class by generating synthetic data proportionate to the majority class [22]. This approach has gained immense popularity due to its simple implementation and remarkable efficacy in various applications [23]. Various studies have introduced enhancements to the original SMOTE algorithm in order to address limitations in its earlier versions. Presently, the literature reports the existence of 85 SMOTE variants to address the class imbalance in machine learning [24]. Further investigations by [25] involved comparing three SMOTE variations, namely Borderline-SMOTE, SVM-SMOTE, and Kmeans-SMOTE, in conjunction with diverse classification algorithms such as Logistic Regression, Random Forest, SVM, and Adaboost. The study demonstrated that Borderline-SMOTE, an extension of SMOTE that generates synthetic minority class samples near the decision boundary, achieved superior evaluation outcomes when combined with the Random Forest algorithm. Random Undersampling (RUS) is another technique used to address data imbalance by reducing the number of instances in the majority class to achieve a balanced dataset. According to [26], RUS significantly enhanced the performance of several classifiers in detecting web attacks using the dataset from CSE-CIC-IDS2018. In addition to oversampling and undersampling, several studies have indicated that data cleaning methods like Tomek Links can enhance classifier performance, as reported in a study by [27]. Furthermore, combining Tomek Links with RUS and SMOTE has shown superior performance compared to using resampling techniques alone, across various classification algorithms

Based on the literature review, both machine learning algorithms such as RF and ANN, as well as resampling approaches show significant promise in addressing classification problems with imbalanced data. To the best of our knowledge, no direct comparison of SNP identification models with these methods has been done. In order to close this gap, this study compares SNP identification models that use ANN and RF and assesses how well machine learning algorithms work when combined with resampling approaches to identify the best accurate method for SNP identification. The paper is structured as follows: the Introduction gives a summary of the issue the study addresses and establishes the context for discussing the difficulties in SNP identification and the need for improved techniques. The dataset used, together with its properties, is described in depth in the Materials and Methods section. It also describes the methods and experimental setups used in the research. The findings are presented and interpreted in the Results and Discussion section, which also discusses their importance and how they advance the field by connecting them to existing research. Finally, the Conclusion summarizes the key insights of the research to present the implications and suggesting potential directions for future investigation.

2. RESEARCH METHOD

2.1. Dataset Description

The study used Single Nucleotide Polymorphisms (SNPs) candidate data from the soybean genome, sourced from previous research [28], and the initial dataset contained 24 features as detailed by [29]. However, research by [30] demonstrated that using just five selected features for SNP calling improved the classification model's F-Measure and significantly reduced computational time compared to using all 24 features. Therefore, this study exclusively employs those five selected features, which comprise 20 chromosomes labeled as Glycine max, chromosome 1 (Gm01) to Gm20 and belong to two classes, namely positive and negative SNPs. The dataset comprises 2,823,602 positive SNPs and 36,631,026 negative SNPs, revealing a significant imbalance, with the negative class having 13 times more instances than the positive class. Figure 1 shows that chromosome Gm11, with a total of 1,653,668 candidate SNPs, has the highest imbalance ratio of 17.9. Therefore, the Gm11 data was used to train the model. Contrarily, chromosome Gm16 which has the fewest candidate SNPs of 1,524,574 was used as the testing data.

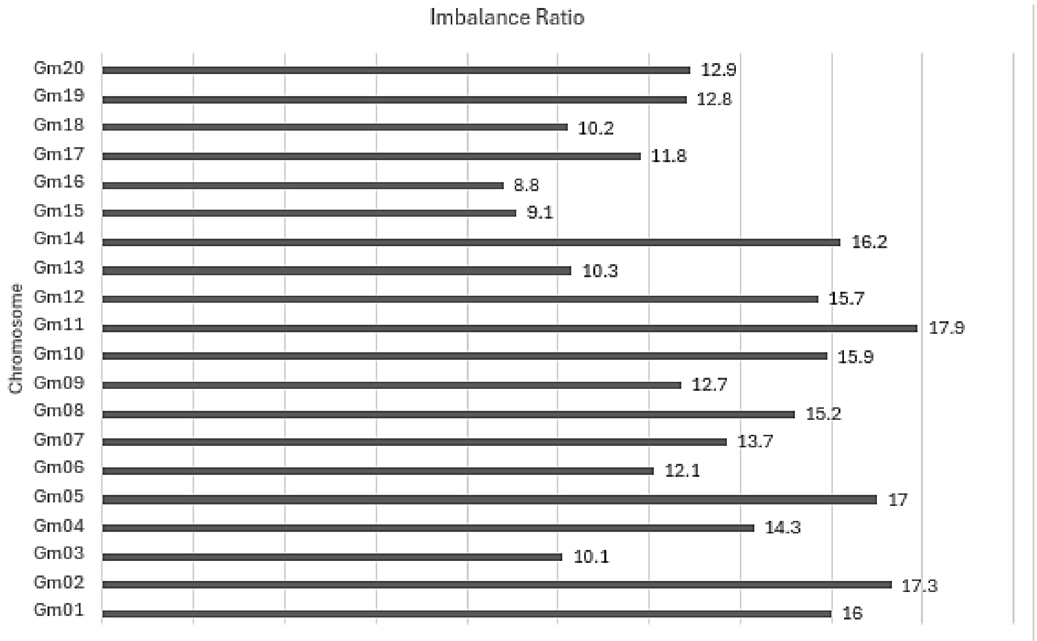

Figure 1. Illustration of imbalance ratio for each chromosome in the soybean genome

2.2. Methods

In general, the workflow of this research can be described through Figure 2. Firstly, the min-max normalization technique transforms each attribute value into a range between 0 and 1 to facilitate uniformity in the dataset. Subsequently, during the resampling stage, the performance of three distinct methods, namely Borderline-SMOTE, Random Undersampling, and Tomek Links, will be evaluated regarding their effectiveness in tackling the challenge of imbalance problem. The resulting balanced dataset will be used to train and test classifiers built using Random Forest (RF) and Artificial Neural Networks (ANN), with hyperparameter tuning applied to optimize performance. This study compares various evaluation metrics, including F-measure, G-Mean, precision, and recall, under different conditions: imbalanced data, oversampled data, undersampled data, and data processed with Tomek Links.

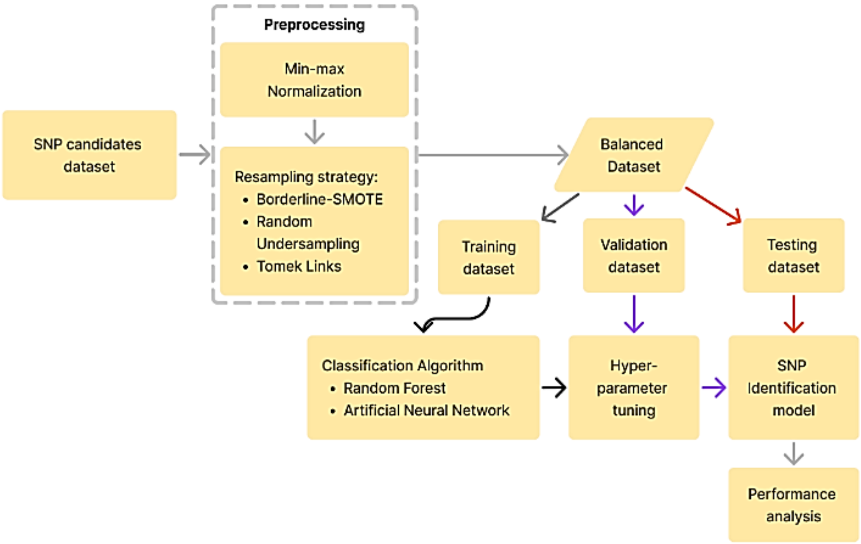

Figure 2. Research workflow to analyze the best approach to identify SNPs.

2.3. Resampling

In this research, we investigate what method is most appropriate to solve imbalanced problems in the context of SNP identification. This research compares three different resampling techniques including Borderline-SMOTE, Random Undersampling (RUS), and Tomek Link. Borderline-SMOTE, an extension of SMOTE, targets creating artificial instances for the minority class near the decision boundary, termed borderline samples. The process divides the dataset into two classes: the minor class (SNP positive) and the major class (SNP negative). In each p_i (where i ranges from 1 to p_{num}) within the minor class P, count the nearest neighbors m in T dataset. All samples from the minor class that are in the nearest neighbor range m are denoted by m' ($0 \leq m' \leq m$). Then the candidate samples from minor class are categorized into three sort of range, namely 'SAFE', 'NOISE', dan 'DANGER' based on equation 1-3.

$$\text{if } m' = m \text{ then } p_i \text{ is "NOISE"} \tag{1}$$

$$\text{if } \frac{m}{2} \leq m' < m \text{ then } p_i \text{ is "DANGER"} \tag{2}$$

$$\text{if } 0 \leq m' < \frac{m}{2} \text{ then } p_i \text{ is "SAFE"} \tag{3}$$

Instances that are in the DANGER range are data that are on the borderline in the P minor class that will be synthesized. This study uses the Borderline-SMOTE Python module from the Imblearn library to handle imbalanced datasets. This module has several input parameters, including sampling_strategy, which is used to adjust the imbalance ratio rate, and k_neighbors, which determines the number of k-nearest neighbors to be sought [31]. The procedure for generating synthetic samples using Borderline-SMOTE is explained in Figure 3.

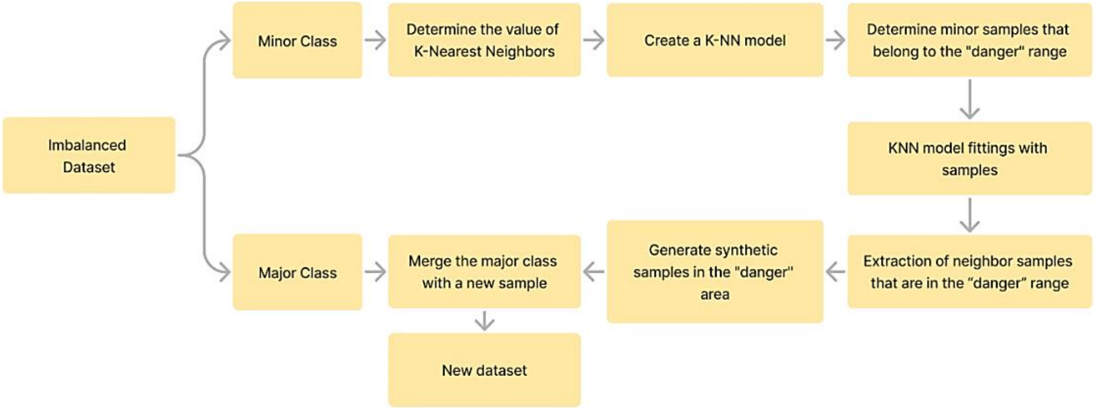
Imbalanced Dataset
Minor Class
Determine the value of K-Nearest Neighbors
Create a K-NN model
Determine minor samples that belong to the "danger" range
KNN model fittings with samples
Extraction of neighbor samples that are in the "danger" range
Generate synthetic samples in the "danger" area
Merge the major class with a new sample
Major Class
New dataset

Figure 3. The oversampling procedure in Borderline-SMOTE

This study explored parameters in the Borderline-SMOTE technique for optimal synthetic data generation to train an SNP identification model. The sampling strategy parameter, representing the ratio of synthetic data generated, was investigated at levels of 0.2, 0.4, 0.6, 0.8, and 1, indicating the percentage of minor data to be generated. Results showed that a sampling strategy of 1 achieved the highest F-measure value, leading to a balanced 1:1 ratio between minor and major classes. Additionally, the k_neighbors' parameter, set to 5, determined the nearest neighbors count for each minor class instance to create synthetic samples.

The opposite of oversampling is undersampling, which involves reducing the number of majority class instances to achieve a more balanced class distribution. Popular undersampling techniques include Random Undersampling and Tomek Link, as recommended in the following reference studies [32], [33]. The advantages of RUS include reducing computational complexity and effectively handling imbalanced data. Nevertheless, this strategy also has its drawback that it has the potential loss of valuable information, as it discards potentially relevant information by removing numerous samples of the majority class [29]. The procedures of RUS can be observed in Figure 4. In this study, several values of sampling strategy were explored, which are 0.2, 0.4, 0.6, 0.8, and 1. These values determine the amount of data that will be eliminated from major class. Based on the experimental results, the highest outcomes were achieved by setting the sampling strategy to 0.4. Therefore, for the subsequent RUS approach, a sampling strategy with a value of 0.4 will be employed.

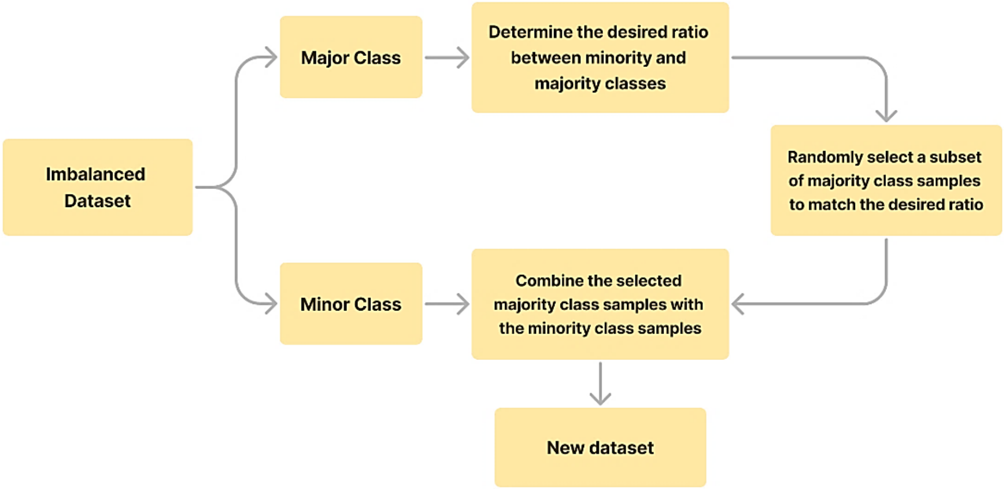
Imbalanced Dataset
Major Class
Determine the desired ratio between minority and majority classes
Randomly select a subset of majority class samples to match the desired ratio
Minor Class
Combine the selected majority class samples with the minority class samples
New dataset

Figure 4. The undersampling procedure in RUS

The Tomek Link algorithm is a data cleansing technique used to handle class imbalance by removing instances. It eliminates pairs of samples from distinct classes that are the nearest neighbors of each other with minimal distance. The procedure involves determining minority and majority class samples and removing samples that form Tomek Links, indicating a clear boundary between classes. By removing Tomek Links, the algorithm enhances class separation and removes noisy or ambiguous samples. The resulting dataset is then used for model training and evaluation to promote better class balance [34]. The procedure of Tomek Link can be seen in Figure 5.

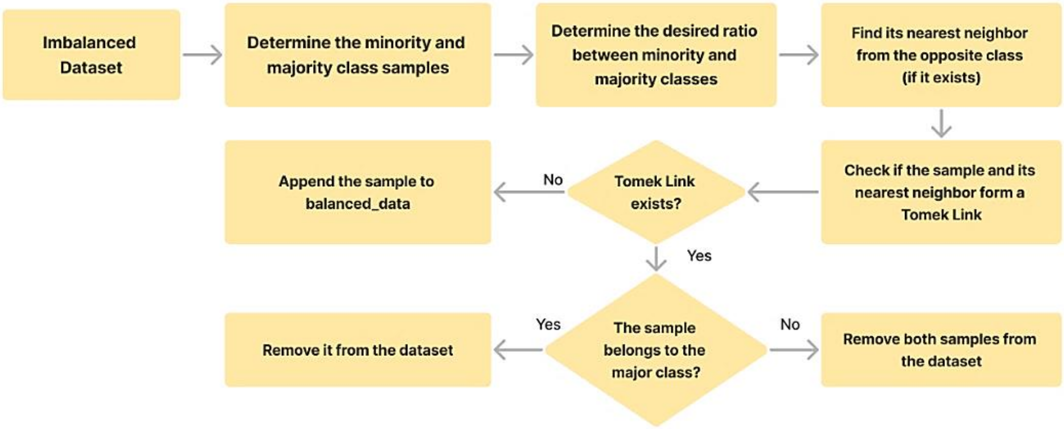
Imbalanced Dataset
Determine the minority and majority class samples
Determine the desired ratio between minority and majority classes
Find its nearest neighbor from the opposite class (if it exists)
Check if the sample and its nearest neighbor form a Tomek Link
Tomek Link exists?
No
Append the sample to balanced_data
Yes
The sample belongs to the major class?
Yes
Remove it from the dataset
No
Remove both samples from the dataset

Figure 5. The data cleansing procedure with Tomek Link

2.4. Hyperparameter tuning for Random Forest (RF)

RF employs a divide-and-conquer strategy to bolster classifier performance. It combines weak learners to form a strong learner, thereby surpassing the performance of individual classifiers [35]. RF mitigates overfitting by constructing multiple decision trees independently, each trained on a different subset of samples, and subsequently combining their outputs to yield robust predictions [36]. Its ensemble nature is particularly effective for handling imbalanced data. The voting mechanism of RF further reduces the impact of misclassifications on the minority class by considering predictions from various decision trees.

In this investigation, hyperparameter tuning for the RF algorithm involved exploring various combinations of parameters, including n_estimators ranging from 10 to 100, to optimize the number of trees. The criterion parameter determined the split type at each node, set to "gini" based on prior studies, as it demonstrated similar performance to entropy but with faster computation [37]. The random state, set to 42 to ensure consistent results across dataset changes [38]. The max_depth parameter, tested at values of 2, 5, 8, 10, and 20, controlled tree depth, with a max_depth of 5 yielding optimal results based on F-measure and Geometric-mean evaluations. The best F-measure value of 91.638, along with the highest G-mean, was achieved with 80 trees, max_depth set to 5, criterion "gini," and random_state 42. Given that 80 trees are faster to build than 90 trees with nearly identical performance, 80 trees were selected for the final model. Hyperparameters were optimized using Grid Search Cross Validation to enhance data prediction accuracy.

2.5. Hyperparameter tuning for Artificial Neural Network

The ANN algorithm gains inspiration from the functioning of the human brain. A “neuron”, the basic computational unit in an ANN, processes and transmits information by receiving input signals from other neurons, each with distinct weights and biases. The model processes data sequentially through interconnected layers, calculates predictive errors at the output layer, and adjusts neuron weights and biases by backpropagating these errors to improve the network's performance. The ANN may have single or multiple hidden layers, thereby improving its capacity for intricate knowledge acquisition and processing [39]. Several key parameters of ANN were optimized in this study, including the maximum number of epochs for training iterations, batch sizes of 64, 128, and 256 to determine the optimal number of observations processed before weight updates, and the selection of the AdamW optimizer with a learning rate of 0.001 for enhanced training performance and stability, consistent with default settings for the Adam optimizer as detailed in references [40].

In this study, two activation functions were compared: softmax and log softmax. Results indicate that log softmax, combined with other parameters, outperforms softmax in terms of F-Measure, consistent with prior findings [41]. Another parameter is the criterion (also known as a loss function) that measures how well the model's predictions match the actual target values in the training data. The goal of training a neural network is to minimize this criterion, which effectively means reducing the discrepancy between predicted and actual values. In this study, we adopt the negative log likelihood criterion as suggested by [42]. Their findings reveal that employing the negative log-likelihood loss significantly enhances prediction rules, ensuring improved calibration and minimal deviation in predicted survival probability. Likelihood-based training also outperforms cross-entropy-based models, yielding noteworthy reductions in prediction errors. The model's performance was assessed using F-measure and Geometric-mean. Among the experiments, three achieved an F-measure of 73 and a G-mean of 93. Notably, the experiment using the log softmax activation function with a batch size of 64 resulted in significantly faster training time. This finding is consistent with [43], which highlighted the superiority of log softmax in higher-dimensional scenarios compared to other activation functions. As a result, these parameters were selected to optimize the model in this study.

3. RESULTS AND ANALYSIS

3.1. Classification with Random Forest and Artificial Neural Networks

This study compares RF and ANN performance in identifying actual SNPs using precision and recall metrics. Balancing recall and precision are challenging, as increasing one often decreases the other. The F-measure effectively combines these aspects for comprehensive evaluation. Figure 6 illustrates that with imbalanced data, ANN outperforms RF in F-measure, scoring 79.75 compared to RF's 69.72. Moreover, ANN achieves nearly balanced precision and recall values, with precision at 80.71 and recall at 78.82. In contrast, RF shows disparate precision and recall values of 85.91 and 58.67, resulting in an F-measure of 69.72. This result observe that ANN maintains a better balance between precision and recall with unbalanced data that minimize false positives and negatives.

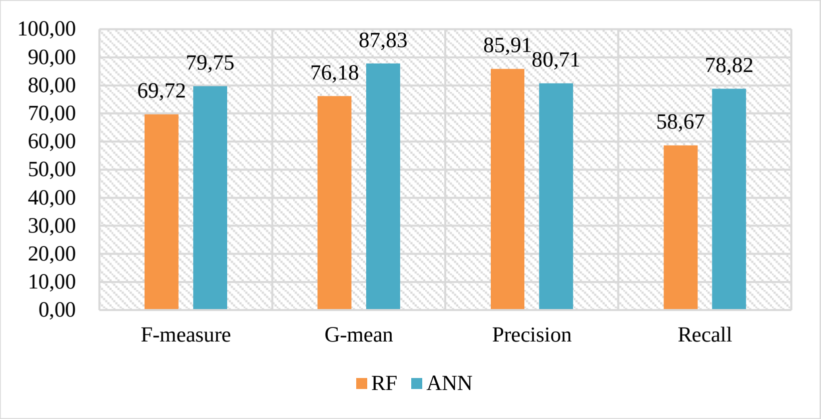

Figure 6. Comparison of RF and ANN in SNP identification model with unbalanced data

Other evaluation metrics commonly used in classification are sensitivity and specificity. Sensitivity measures the model's performance in identifying positive or minority class samples, while specificity assesses its ability to detect negative or majority class samples that presents a trade-off. The geometric mean (G-Mean) evaluates classification balance across classes, with low G-Mean indicating poor positive case classification despite accurate negative case classification. Based on Figure 6, without dataset resampling, ANN also exceeds RF in G-Mean with 87.83 whereas RF only obtains 76.18. This improvement can be attributed to ANN’s robustness in learning from imbalanced data, where other methods might struggle. According to theoretical frameworks, ANN's capability to adjust weights dynamically during training allows for better handling of class imbalance, though it may still favor the negative class as the imbalance ratio increases [43].

The implication of this findings are considerable for genomics research and related fields. ANN’s ability to deal with unbalanced datasets makes it a valuable tool for applications where data imbalance is prevalent, such as medical diagnosis, fraud detection, and anomaly detection. This capability can lead to more accurate predictions and better decission-making in critical areas. Moreover, the dynamic weight adjustment capability of ANN during training allows for better handling of class imbalance, though it may still favor the negative class as the imbalance ratio increases.

3.2. The Influence of Resampling Strategy on Classification Using RF and ANN

To determine the optimal resample combination to RF in identifying valid SNPs, several experimental scenarios were conducted, and the outcomes are depicted in the following Figure 7.

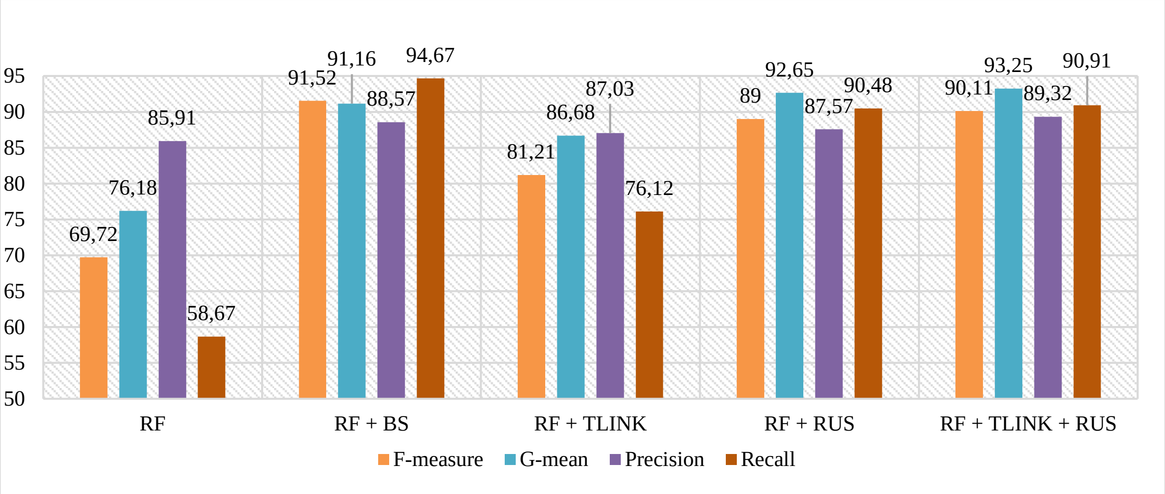

Figure 7. Evaluation of RF combined with Borderline_SMOTE, Tomek Link and RUS.

The use of Borderline-SMOTE within RF significantly increased F-Measure from 69.72 to 91.52. Similarly, integrating RUS and T-Link raised F-Measure to 89 and 81.21, respectively. The study also combined T-Link with RUS, in line with previous research suggesting T-Link's role as a data cleaning method integrated with RUS. This integration improved F-Measure to 90.11, slightly surpassing RF+RUS and

RF+TLink alone. Furthermore, improvement is also evident in the G-Mean metric, which initially scored only 76.18. After undergoing resampling, its values increased to 91.16 for RF+BS, 86.68 for RF+TLink, and 92.65 for RF+RUS. In addition, the application of T-Link combined with RUS also increases the G-Mean value into the highest score of 93.25. The rise in G-Mean stems from increased recall/sensitivity that represent the classifier's ability to accurately identify positive instances. Resampling techniques notably boosted sensitivity from 58.67 to 94.67 for RF+BS, 90.48 for RF+RUS, and 90.91 for RF+TLink+RUS. Improved sensitivity is crucial in scenarios like SNP identification, where accurate classification of positive instances is essential and indicate the model's enhanced ability to avoid false negatives and identify positive instances.

This study also observes the impact of applying resampling techniques to ANN, with the results of several experiments summarized in the following Figure 8.

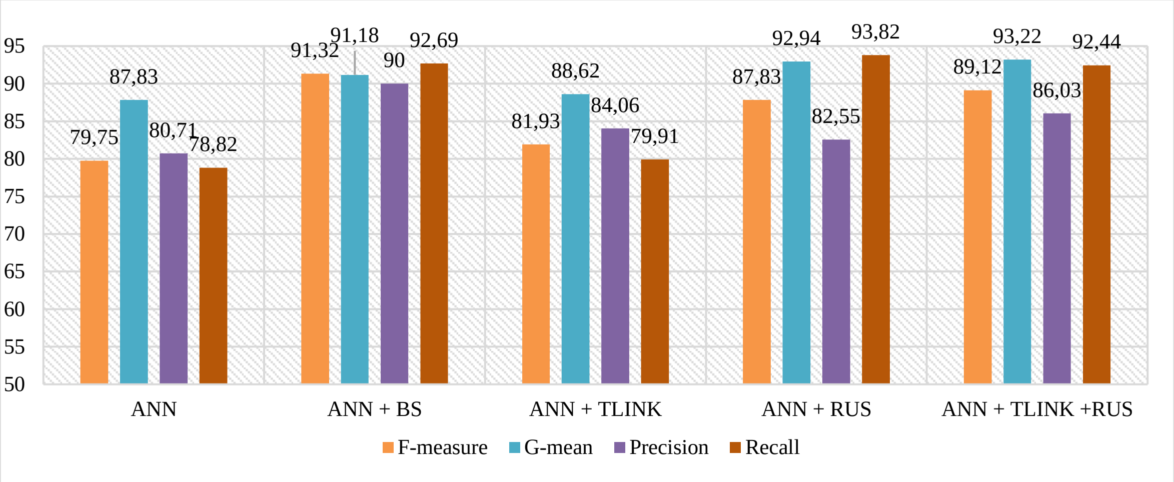

Figure 8. Evaluation of ANN combined with Borderline_SMOTE, Tomek Link and RUS.

Figure 8 demonstrates that employing Borderline-SMOTE significantly enhances ANN performance, consistently achieving scores exceeding 90 across all metrics, with an F-Measure of 91.32. Conversely, using T-Link yields the lowest scores in both RF and ANN models. This emphasizes that solely relying on T-Link may not adequately tackle the issues presented by imbalanced data. In principle, T-Link only removes a small portion of the majority data, specifically those meeting the Tomek-Link criteria as illustrated in Figure 9. Consequently, minimal alteration occurs in the data distribution and class imbalance ratio, resulting in less satisfactory evaluation scores. However, T-Link can serve as a valuable data cleaning tool that could enhance RUS technique performance, and optimizing computational efficiency in developing superior classification models.

We also benchmarked our results against previous studies on SNP discovery conducted by [44] and [45], which used Decision Tree C4.5 and SVM, respectively, with precision (also known as Positive Predictive Value, PPV) as the evaluation metric. Our results show that the ANN-Borderline SMOTE in this study achieved the best precision of 90%, surpassing both previous methods, which achieved precision values of 84.8% and 61%, respectively. This demonstrates the effectiveness of proposed methods especially ANN-Borderline SMOTE in SNP identification. The implications of this research highlight the effectiveness of integrating resampling techniques, particularly Borderline-SMOTE and RUS, in improving classification performance on imbalanced datasets. These findings suggest that RF and ANN models benefit significantly from these methods, particularly in terms of sensitivity and F-Measure. However, one limitation of this study is its focus on only a few resampling techniques (Borderline-SMOTE, RUS, and T-Link) and their application in RF and ANN models. This study does not consider other advanced oversampling or undersampling methods, which may offer additional benefits. Future research could explore a wider range of resampling and data-cleaning techniques, including more hybrid methods, to further optimize model performance on imbalanced datasets. Furthermore, exploring the use of ensemble methods or deep learning techniques in combination with resampling strategies could potentially lead to even more robust classification models.

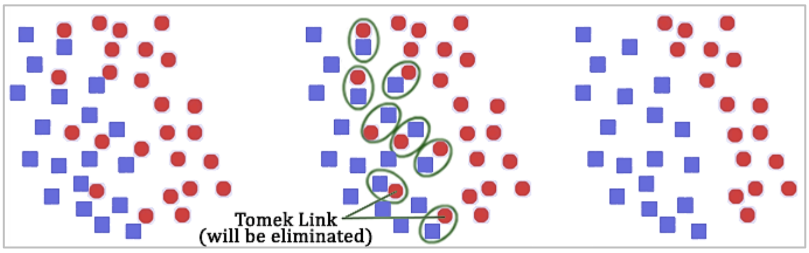

Figure 9. Tomek-Link data cleaning illustration.

4. CONCLUSION

This study evaluates the performance of Random Forest (RF) and Artificial Neural Networks (ANN) in identifying Single Nucleotide Polymorphisms (SNPs) using various metrics, including precision, recall, F-measure, and G-Mean. The analysis reveals that ANN outperforms RF in balancing precision and recall, particularly with imbalanced data. ANN achieves a higher F-measure of 79.75 compared to RF's 69.72. In contrast, RF exhibits a discrepancy between precision (85.91) and recall (58.67), resulting in a lower F-measure. Additionally, ANN demonstrates superior performance in G-Mean, which reflects a better balance between positive and negative class classification.

The study also explores the impact of various resampling techniques on RF and ANN. Borderline-SMOTE significantly enhances RF's performance and improves the F-measure from 69.72 to 91.52. This improvement reflects Borderline-SMOTE's effectiveness in generating synthetic samples to balance class distributions and improve RF and ANN to classify SNPs accurately. Combining resampling methods like Random Undersampling (RUS) and Tomek Link also boosts F-measure and G-Mean scores. It indicates improved model performance. For ANN, Borderline-SMOTE consistently yields high scores across all metrics, with an F-measure of 91.32, while Tomek Link alone shows less effectiveness. Resampling techniques notably increase sensitivity that highlight their crucial role in accurately identifying positive instances in SNP classification. These findings have implications for improving SNP identification in genomic research and provide insightful information on how to optimize machine learning methods for unbalanced datasets. However, this study has limitations, such as focusing on only a few resampling methods and using a dataset limited to SNP candidates from soybean genomes. Future studies should investigate a wider range of resampling techniques or hybrid approaches to improve the efficiency of machine learning models in SNP detection. Significant advances could also result from exploring deep learning systems, particularly those designed to handle highly imbalanced data. To gain a more general understanding, future research should employ diverse genomic datasets to ensure the robustness of the proposed methodology. Further investigation in these areas will help refine and streamline techniques for addressing imbalanced datasets in genomic research, ultimately enhancing our ability to detect genetic variants more effectively.

ACKNOWLEDGEMENTS

The author expresses gratitude to Universitas Sumatera Utara for providing the research grant under the Penelitian Dasar scheme with grant number 1705/UN5.1.R/PPM/2022.

REFERENCES

[1] L. Picoult-Newberg et al., "Mining SNPs from EST databases," Genome Res, vol. 9, no. 2, pp. 167--174, 1999, doi: 10.1101/gr.9.2.167.

[2] P. Nowotny, J. M. Kwon, and A. M. Goate, "SNP analysis to dissect human traits," Curr Opin Neurobiol, vol. 11, no. 5, pp. 637–641, 2001, doi: https://doi.org/10.1016/S0959-4388(00)00261-0.

[3] I. Joshi et al., "15 - Artificial intelligence, big data and machine learning approaches in genome-wide SNP-based prediction for precision medicine and drug discovery," in Big Data Analytics in Chemoinformatics and Bioinformatics, S. C. Basak and M. Vračko, Eds., Elsevier, 2023, pp. 333–357. doi: https://doi.org/10.1016/B978-0-323-85713-0.00021-9.

[4] J. Candotti et al., "Haplotype mining panel for genetic dissection and breeding in Eucalyptus.," Plant J, vol. 113, no. 1, pp. 174—185, 2022. doi: 10.1111/tpj.16026

[5] O. A. Gutiérrez et al., "SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of Theobroma cacao L.," Tree Genet Genomes, vol. 17, no. 3, 2021, doi: 10.1007/s11295-021-01507-w.

[6] J. Xue et al., "An overview of SNP-SNP microhaplotypes in the 26 populations of the 1000 Genomes Project," Int J Legal Med, vol. 136, no. 5, pp. 1211–1226, 2022, doi: 10.1007/s00414-022-02820-2.

[7] L. S. Hasibuan, N. Hudachair, and M. A. Istiadi, "Bootstrap aggregating of classification and regression trees in identification of single nucleotide polymorphisms," in 2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017, 2018. doi: 10.1109/ICACSIS.2017.8355068.

[8] M. Wasikowski and X. Chen, “Combating the Small Sample Class Imbalance Problem Using Feature Selection,” IEEE Trans Knowl Data Eng, vol. 22, no. 10, pp. 1388–1400, 2010, doi: 10.1109/TKDE.2009.187.

[9] M. Koziarski, “Potential Anchoring for imbalanced data classification,” Pattern Recognit, vol. 120, 2021, doi: 10.1016/j.patcog.2021.108114.

[10] A. S. More and D. P. Rana, “Review of random forest classification techniques to resolve data imbalance,” in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 72–78. doi: 10.1109/ICISIM.2017.8122151.

[11] A. C. Neocleous, K. H. Nicolaides, and C. N. Schizas, “Intelligent Noninvasive Diagnosis of Aneuploidy: Raw Values and Highly Imbalanced Dataset,” IEEE J Biomed Health Inform, vol. 21, no. 5, 2017, doi: 10.1109/JBHI.2016.2608859.

[12] S.-H. Oh, “A Statistical Perspective of Neural Networks for Imbalanced Data Problems,” International Journal of Contents, vol. 7, no. 3, 2011, doi: 10.5392/ijoc.2011.7.3.001.

[13] L. Breiman, “Random forests,” Mach Learn, vol. 45, pp. 5–32, 2001.

[14] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” BMC Med Inform Decis Mak, vol. 11, no. 1, 2011, doi: 10.1186/1472-6947-11-51.

[15] S. Sakr et al., “Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry Ford exercIse testing (FIT) project,” BMC Med Inform Decis Mak, vol. 17, no. 1, 2017, doi: 10.1186/s12911-017-0566-6.

[16] J. H. Ma, Z. Feng, J. Y. Wu, Y. Zhang, and W. Di, “Learning from imbalanced fetal outcomes of systemic lupus erythematosus in artificial neural networks,” BMC Med Inform Decis Mak, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01486-x.

[17] S. Bagui and K. Li, “Resampling imbalanced data for network intrusion detection datasets,” J Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00390-x.

[18] R. Taghizadeh-Mehrjardi et al., “Synthetic resampling strategies and machine learning for digital soil mapping in Iran,” Eur J Soil Sci, vol. 71, no. 3, 2020, doi: 10.1111/ejss.12893.

[19] C. Zhang, P. Soda, J. Bi, G. Fan, G. Almpanidis, and S. Garcia, “An Empirical Study on the Joint Impact of Feature Selection and Data Re-sampling on Imbalance Classification,” Appl Intell, vol. 53, no. 5, pp. 5449—5461, 2023, doi: https://doi.org/10.1007/s10489-022-03772-1.

[20] W. A. Kusuma, A. S. Rahmi, and R. Heryanto, “Implementation of hybrid sampling technique for predicting active compound and protein interaction in unbalanced dataset,” in IOP Conference Series: Earth and Environmental Science, vol. 335, no. 1, pp. 012005, 2019. doi: 10.1088/1755-1315/335/1/012005.

[21] I. Sadgali, N. Sael, and F. Benabbou, “Bidirectional gated recurrent unit for improving classification in credit card fraud detection,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, 2021, doi: 10.11591/ijeecs.v21.i3.pp1704-1712.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[23] D. Elreedy and A. F. Atiya, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” Inf Sci (N Y), vol. 505, 2019, doi: 10.1016/j.ins.2019.07.070.

[24] G. Kovács, “Smote-variants: A python implementation of 85 minority oversampling techniques,” Neurocomputing, vol. 366, 2019, doi: 10.1016/j.neucom.2019.06.100.

[25] X. Zheng, “SMOTE Variants for Imbalanced Binary Classification: Heart Disease Prediction,” J Chem Inf Model, vol. 21, no. 1, 2020.

[26] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, “Investigating rarity in web attacks with ensemble learners,” J Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00462-6.

[27] E. AT, A. M, A.-M. F, and S. M, “Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method,” Global Journal of Technology and Optimization, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.

[28] M. A. Istiadi, W. A. Kusuma, and I. M. Tasma, “Application of decision tree classifier for single nucleotide polymorphism discovery from next-generation sequencing data,” in Proceedings - ICACSIS 2014: 2014 International Conference on Advanced Computer Science and Information Systems, 2014. doi: 10.1109/ICACSIS.2014.7065832.

[29] L. Sahrina Hasibuan, S. Nabila, N. Hudachair, and M. Abrar Istiadi, “Evaluation of F-Measure and Feature Analysis of C5.0 Implementation on Single Nucleotide Polymorphism Calling,” Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), vol. 1, no. 1, pp. 1–5, 2018.

[30] R. Nurhasanah, A. Buono, and W. A. Kusuma, “COMBINING SIGNAL TO NOISE RATIO AND UNDERSAMPLING IN SINGLE NUCLEOTIDE POLYMORPHISMS IDENTIFICATION,” Indian Journal of Computer Science and Engineering, vol. 14, no. 3, pp. 490–499, Jun. 2023, doi: 10.21817/indjcse/2023/v14i3/231403029.

[31] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” Journal of Machine Learning Research, vol. 18, 2017.

[32] A. Thumpati and Y. Zhang, “Towards Optimizing Performance of Machine Learning Algorithms on Unbalanced Dataset,” 2023. doi: 10.5121/csit.2023.131914.

[33] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, “Detecting web attacks using random undersampling and ensemble learners,” J Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00460-8.

[34] M. U. Khan, S. U. J. Lee, S. Abbas, A. Abbas, and A. K. Bashir, “Detecting Wake Lock Leaks in Android Apps Using Machine Learning,” IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3110244.

[35] G. Kawamura, S. Seno, Y. Takenaka, and H. Matsuda, “A Combination Method of the Tanimoto Coefficient and Proximity Measure of Random Forest for Compound Activity Prediction,” IPSJ Digital Courier, vol. 4, 2008, doi: 10.2197/ipsjdc.4.238.

[36] A. S. More and D. P. Rana, “Performance enrichment through parameter tuning of random forest classification for imbalanced data applications,” Mater Today Proc, 2022, doi: 10.1016/j.matpr.2021.12.020.

[37] L. Roberts, L. Razoumov, L. Su, and Y. Wang, “Gini-regularized Optimal Transport with an Application to Spatio-Temporal Forecasting,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.02512

[38] R. P. Pratama and W. Maharani, “Predicting Big Five Personality Traits Based on Twitter User U sing Random Forest Method*,” in 2021 International Conference on Data Science and Its Applications, ICoDSA 2021, 2021. doi: 10.1109/ICoDSA53588.2021.9617501.

[39] J. W. Huang, C. W. Chiang, and J. W. Chang, “Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise,” Eng Appl Artif Intell, vol. 75, 2018, doi: 10.1016/j.engappai.2018.07.010.

[40] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in 7th International Conference on Learning Representations, ICLR 2019, International Conference on Learning Representations, ICLR, 2019.

[41] Ü. Yllmaz, C. Gezer, Z. Aydln, and V. C. Güngör, “Data Mining Techniques in Direct Marketing on Imbalanced Data using Tomek Link Combined with Random Under-sampling,” in ACM International Conference Proceeding Series, 2021. doi: 10.1145/3471287.3471299.

[42] S. G. Zadeh and M. Schmid, “Bias in Cross-Entropy-Based Training of Deep Survival Networks,” IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 9, pp. 3126–3137, Sep. 2021, doi: 10.1109/TPAMI.2020.2979450.

[43] Z. ao Huang, Y. Sang, Y. Sun, and J. Lv, “A neural network learning algorithm for highly imbalanced data classification,” Inf Sci (N Y), vol. 612, 2022, doi: 10.1016/j.ins.2022.08.074.

[44] L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, I. Y. Choi, P. B. Cregan, and C. P. Van Tassell, “Application of machine learning in SNP discovery,” BMC Bioinformatics, 2006, doi: 10.1186/1471-2105-7-4.

[45] L. S. Hasibuan, W. A. Kusuma, and W. B. Suwamo, “Identification of single nucleotide polymorphism using support vector machine on imbalanced data,” Proceedings - ICACSIS 2014: 2014 International Conference on Advanced Computer Science and Information Systems, no. June, pp. 375–379, 2014, doi: 10.1109/ICACSIS.2014.7065854.

BIBLIOGRAPHY OF AUTHORS

Rossy Nurhasanah, received her Master of Computer Science from IPB University, Bogor, Indonesia, in 2015. She is an Assistant Professor at the Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia. She is also a member of the Indonesian Society of Bioinformatics and Biodiversity. Her research interests include bioinformatics, artificial intelligence, computer vision and object detection.

Dedy Arisandi, completed his Master's degree in Information Technology at Universitas Putra Indonesia in 2009. He is currently the Head of the Undergraduate Program in Information Technology at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara (USU). His research areas include Intelligent Systems, Computer Vision, and Machine Learning.

Fanindia Purnamasari, earned her Master's degree from Universiti Kebangsaan Malaysia in 2017. She joined Universitas Sumatera Utara as a lecturer in 2019 and her research interests include human-computer interaction, data science, and artificial intelligence.

Hayatunnufus holds a master’s degree in computer science from Universitas Gadjah Mada (UGM). She is a lecturer in the undergraduate Computer Science program at Universitas Sumatera Utara (USU), with research interests in IoT, AIoT, and data science.

Daisy Sere Damara Simangunsong, a graduate of the Information Technology undergraduate program at Universitas Sumatera Utara, with a specialization in Machine Learning. She currently works as a Data Engineer at PT. Bank Rakyat Indonesia (Persero).

Aflah Mutsanni Pulungan is a graduate of the Information Technology undergraduate program at Universitas Sumatera Utara, specializing in Machine Learning.