COMBINING SIGNAL TO NOISE RATIO AND UNDERSAMPLING IN SINGLE NUCLEOTIDE POLYMORPHISMS IDENTIFICATION

Rossy Nurhasanah

Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia
rossynurhasanah@usu.ac.id

Agus Buono

Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor 16680, Indonesia
agusbuono@apps.ipb.ac.id

Wisnu Ananta Kusuma*

Tropical Biopharmaca Research Center, IPB University, Bogor, 16154, Indonesia
Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor 16680, Indonesia
*Corresponding author: ananta@apps.ipb.ac.id

Abstract

Imbalanced data distribution is a challenge in identifying Single Nucleotide Polymorphisms (SNP) as the amount of false SNP data is far greater than actual SNP data, leading to inappropriate classification results. To overcome this issue, we propose using the Signal to Noise Ratio as a feature selection approach combined with an undersampling technique. We recommend the five best features out of the 24 available ones: maximum quality of minor alleles, average quality of minor alleles, minor allele frequencies, probability of error, and balance of alleles. Our proposed model, which applies five selected features followed by the undersampling process, achieves the highest average sensitivity and F-Measure of 0.96 and 0.92, respectively, while also improving computation speed by up to 28 times. Our strategy is especially suitable for classifications with an imbalanced data distribution, particularly for large data sizes.

***Keywords*: feature selection; undersampling; imbalanced data; single nucleotide polymorphisms**

1. Introduction

In the last few years, molecular markers have proven to be effective tools in the identification of genes that determine various crop properties with economic value [Jannink *et al*. (2010); Lema (2018); Wani *et al*. (2018)]. One of the most used molecular markers is Single Nucleotide Polymorphism (SNP) [Yang *et al*. (2019)]. SNPs are DNA markers that indicate variations in a genome sequence at a single position caused by a change of one nucleotide among individuals [Gupta *et al.* (2001)]. Each organism contains many SNPs that form unique patterns and have biological implications that contribute to species diversity [Kwok and Chen (2003); Liao and Lee (2010)]. These variations are responsible for most trait differences within a species. Typically, SNP identification involves comparing the aligned reads with the reference sequence or the sequence of assembled consensus reads with the reference sequence. The SNP locations are identified by detecting the mismatches between the aligned genomes and the reference sequence [Han *et al.* (2020)]. Some SNPs found in the human genome have been used to diagnose diseases, while in agriculture, particularly in plant breeding, researchers utilize SNP markers to obtain superior varieties that can produce high-quality crops and resist multiple environmental conditions [Gupta *et al.* (2001)].

Identifying accurate SNPs remains a challenge due to the high error rate of data generated from Next Generation Sequencers (NGS) [Korani, Clevenger, Chu, & Ozias-Akins, 2019; Nurhasanah, Hasibuan, & Kusuma, 2020]. These errors may stem from base-calling or the alignment process, making it difficult to use NGS data for SNP mining [Istiadi *et al.* (2014)]. As a result, many of the single base variations in DNA sequences originate from errors rather than actual SNPs, resulting in imbalanced class problems where the number of actual SNPs, the most important class for classification, is significantly lower than that of false SNPs. This imbalance can result in biased machine learning models that are less capable of detecting minor classes [Wang *et al*. (2022)], which are more crucial for classification purposes [Soufan *et al.* (2018)]; [Koziarski, (2021)]; [Esposito *et al.* (2021)]. Furthermore, the large size of NGS data leads to complex and time-consuming computations [Li *et al.* (2016)]; [Ko *et al.* (2018)].

Several studies have aimed to improve the accuracy of SNP identification [Kirov *et al.* (2016); Korani *et al.* (2019)]. Research conducted by Matukumalli *et al.* (2006) developed a SNP classifier using a decision tree algorithm on six types of soybean genome data. The study selected relevant features that influenced the polymorphism scoring decision, resulting in the recommendation of 16 optimized feature sets and an improved positive predictive value of 84.8%. Another model, SNPSVM, used the Caucasus female exome sample from Eastern Europe region and selected features that achieved high sensitivity and specificity. This model used two strategies: initializing three low-performing attributes and sequentially adding new features, and a leave-one-out procedure that calculated each feature's impact on the model's performance. SNPSVM achieved a sensitivity level of 95% using the remaining 15 features [O'Fallon *et al.* (2013)].

To address the issue of imbalanced class distribution, a commonly used method is to apply resampling techniques such as undersampling the majority class or over-sampling the minority class. [Istiadi *et al.* (2014)] built a classifier to identify SNP in soybean genome using C4.5 decision tree algorithm following the pipeline proposed by O'Fallon *et al.* (2013). The researchers extracted 24 features of SNPs based on a recommendation of Matukumalli *et al.* (2006). However, their model performed poorly with low sensitivity (56.7%) due to imbalanced data, but it was improved after conducting a simple random undersampling.

Using feature selection technique is another strategy for addressing the issue of class imbalance. In addition to its benefit in reducing dimensions, the significance of feature selection in tackling the problem of imbalance has been widely explored in recent studies. Research by Wasikowski and Chen (2010) investigated the effectiveness of different feature selection methods, including Pearson correlation coefficient, $\chi 2$, and Signal to Noise Ratio (SNR) in unbalanced data distribution. The outcome of the study showed that the SNR correlation coefficient was among the most effective techniques in comparison to other feature selection methods for classifying data with imbalanced distributions.

SNR is a technique that is frequently employed to rank features and assess the significance of genes in gene expression analysis. Previous studies have utilized SNR to differentiate between two classes by determining the maximum distance between the mean values and the minimum variation of expressions within each class [Cuperlovic-Culf *et al.* (2005)]. In a study that aimed to select biomarkers in Leukemia datasets, genes with similar expressions were grouped using K-Means Clustering, followed by selecting the best feature for each cluster based on SNR score. This approach yielded the best outcome when applied in Support Vector Machine [Mishra and Sahu (2011a)]. Another study compared SNR with t-statistics methods to identify the most relevant genes using 65 genes and 14 samples from each dataset. The results showed that SNR was more effective than t-statistics and achieved 100% accuracy in all datasets except breast cancer, where 5 genes were selected [Mishra and Sahu (2011b)]. The findings indicate that SNR is a reliable approach for evaluating the discriminative power of genes. A similar study also concluded that SNR analysis is useful in selecting the most important genes and reducing the data dimension simultaneously. The study found that using Probabilistic Neural Networks with SNR achieved the best performance for various types of cancer classifications within a reasonable amount of time, and the SNR analysis provided several potential genes for further research [Huang and Liao (2003)].

Previous studies have shown that separately, each feature selection and resampling implementation has been proven to improve classification performance, especially in imbalanced data distribution. However, some researchers suggest that combining these two approaches could provide an optimal classifier model while reducing computational requirements [Haury *et al.* (2011)]. The combination of feature selection with resampling is also expected to yield a more effective and efficient method for overcoming classification problems with unbalanced data [Zhang *et al*. (2022)]. Although this approach has not been applied to SNP identification, this work aims to investigate the impact of combining feature selection with the resampling technique in this context.

In previous studies on SNP identification, different features were used, but some features were also found to overlap [Matukumalli *et al.* (2006); O'Fallon *et al.* (2013); Lam *et al.* (2010)]. In this research, we utilized the 24 features extracted by prior studies [Istiadi *et al.* (2014)] to construct an SNP identification model. Initially, we used SNR as a feature selection technique to identify the most important and relevant features in SNP identification. After feature selection, the dataset was balanced by employing the random undersampling method. Finally, we trained the Support Vector Machine (SVM) classifier using the final dataset and evaluated its efficiency and effectiveness based on various metrics.

2. Material and Method

2.1. *Data Explanation*

The genomic data used in this study were derived from a previous study [Lam *et al.* (2010)]. The whole-genome data used was 955.1 Mb base, of which 937.3 Mb base was successfully mapped into 20 chromosomes (labeled Gm1 to Gm20). The values of each feature were extracted using the SNPSVM library [O'Fallon *et al.* (2013)] and modified according to the needs by Istiadi *et al.* (2014). Every SNP candidate was represented in 24 features. In the 20 chromosomes, 39,454,648 candidate SNPs were found, of which only 2,823,603 candidates were designated as actual SNPs, while the rest were false SNPs, which originated from errors. An illustration of the comparison of the number of false and actual SNP data on the 20 chromosomes can be seen in Figure 1.

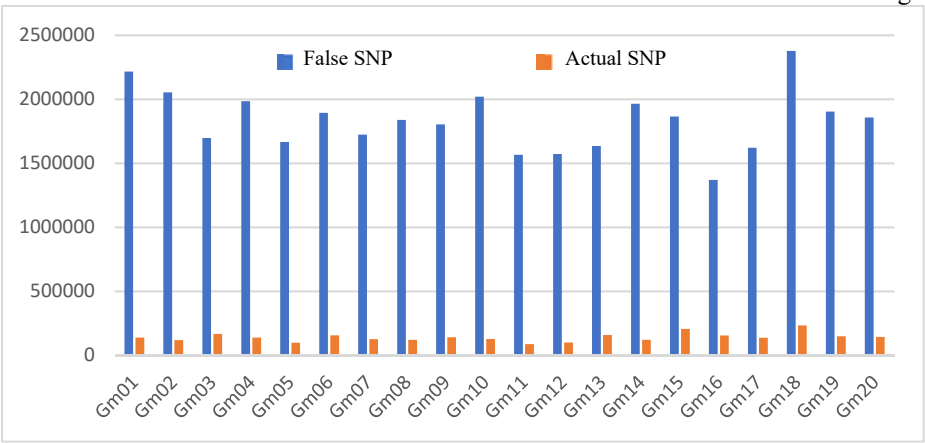

Fig. 1. Comparison of the number of false SNPs with Actual SNPs on each chromosome

From the illustration in Figure 1, there is a significant difference in the amount of data from the two classes. The ratio of false SNP to actual SNP is around 13: 1. Unbalanced data distribution in binary classification can harm the performance of the SNP identifier model because machine learning algorithms generally tend to produce unsatisfactory classifier when trained with an unbalanced data set.

Before being used further, firstly the raw data is normalized using the min-max scaler method. All features' values are normalized using the min-max scaler following Equation (1) so that all feature values are at the same interval, between 0 and 1. The feature selection step utilized the SNR method, and the resampling stage employed a 1:1 undersampling approach that was previously recommended for achieving the best outcomes [Hasibuan *et al.* (2014)]. To classify the data, we used the Support Vector Machine (SVM) algorithm, which is known for its effectiveness in handling imbalanced classification problems, as it only uses support vectors to build the model and ignores most other samples that are not support vectors [Tang *et al.* (2009)]. This leaves SVM to be less affected by negative samples which are far from the hyperplane, even though the number of negative instances is much higher.

$$\text{Newdata} = \frac{\text{X}-\min(\text{X})}{max(X)-min(\text{X})} \qquad (1)$$

2.2. *Feature selection using Signal to Noise Ratio*

The reduction of features in a dataset can bring several benefits such as faster model training, lower susceptibility to overfitting, and reduced storage, memory, and processing requirements during data analysis [Guyon and De (2003)]. Feature selection involves identifying and selecting the most appropriate subset of features from an original dataset [Ditzler and Polikar (2013)]. The main objectives of feature selection are to identify the most influential features for classification and eliminate irrelevant features, thereby simplifying the classification

process and speeding up the computational time by reducing the input dimensions. Additionally, feature selection can improve the classification quality in terms of accuracy since several features may not have any impact on the classification, and in some cases, they may even decrease the performance of the classifier [Kudo and Sklansky (2000)].

The Signal to Noise Ratio (SNR) is a measure of the ratio between the strength of the signal being studied and the noise present in the signal. It is used as a correlation coefficient that compares the ratio of differences between the averages of two classes with the number of standard deviations of the two classes. When there is a significant difference between the averages of the two classes for a feature, the probability of a sample being incorrectly classified is reduced. Conversely, if the average class on a feature is only slightly different, then the probability of a sample being incorrectly classified is increased. The SNR score can be calculated using equation (2) as shown below.

$$\text{SNR score} = \frac{\mu 1 - \mu 2}{\sigma 1 - \sigma 2} \qquad (2)$$

$\mu 1$ and $\mu 2$ are the averages of feature values in the positive SNP class and the negative SNP class, meanwhile, $\sigma 1$ and $\sigma 2$ are the standard deviations of feature values in the positive and negative SNP classes. Few studies have used SNR as a feature selection metric. Research applied SNR to leukemia classification in calculating the correlation between genes with class differentiation [Golub *et al.* (1999)]. Features that obtain a high SNR score indicate a strong correlation between them and their ability to distinguish classes. Features with high SNR values are informative features that can be selected for classification [Gunavathi and Premalatha (2014)]. While features with low SNR scores or approaching 0 represent that these features do not affect classification, and features having high SNR scores or approaching 1 represent that the feature has a strong contribution to the classification [Verikas and Bacauskiene (2002)].

2.3. *SNP Classification*

This research aimed to identify SNP in the soybean genome through a three-step process consisting of feature selection, undersampling, and classification. The workflow for this process is illustrated in Figure 2.

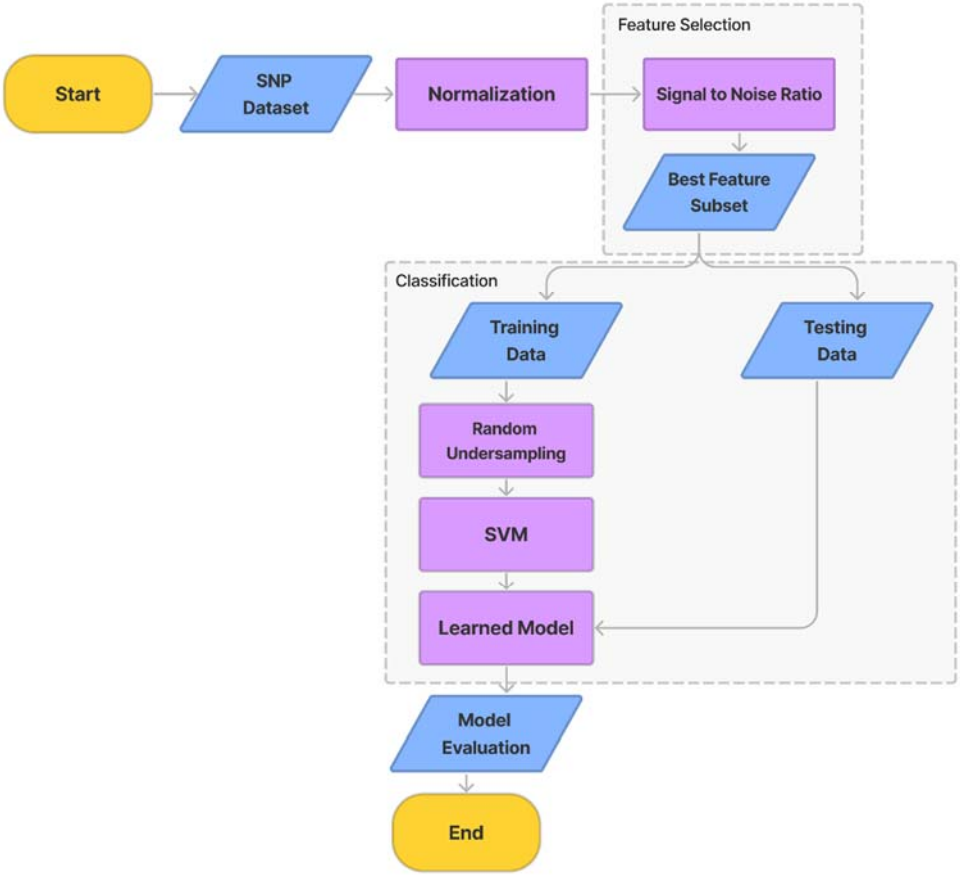

Fig. 2. The workflow of SNP identification using feature selection and undersampling approach.

Compared to other classifiers, SVM is more suitable for facing imbalanced class problems. SVM only involves support vectors in constructing the model, and the remaining samples far away from the hyperplane could be denied without affecting classification performance [Akbani *et al.* (2004)]. However, the SVM performance could be decreased due to a very imbalanced class in which the number of negative samples is significantly higher than the number of positive ones [Tang *et al.* (2009)]. Some recent works have been conducted to improve the classification performance of SVM on imbalanced datasets. However, they could not have addressed efficiency better and could take a longer time for classification than a standard SVM. This becomes another challenge in using SVM since it needs much time to do the computation, especially when working with large datasets. We conducted SVM training and classification procedures using e1071, a free LIBSVM library for R Programming [Meyer *et al.* (2014)]. We used the Radial Basis Function kernel and applied grid search with 10-fold-cross-validation to optimize the C and γ parameters.

2.4. *Evaluation Metrics*

This research investigates the effect of integrating feature selection and undersampling techniques in the identification of SNPs with an SVM classifier. The study concentrates on evaluating the efficacy and efficiency of this approach. Efficiency is characterized by the swiftness of the model in identifying new samples, measured by the time taken to execute the classification. The paper defines effectiveness as the ability of the model to detect positive SNP instances, emphasizing their importance in the classification process.

Classifier performance is typically evaluated using accuracy metrics. However, when dealing with unbalanced data, accuracy can be an inappropriate measure as a classifier may predict the negative class with high accuracy but perform poorly in identifying the minority class. In such cases, alternative metrics are required to assess a classifier's ability to detect positive samples. This study utilizes metrics such as Geometric Mean and F-Measure to evaluate the classifier's performance in identifying the presence of positive samples in class imbalance scenarios [Yu *et al.* (2022)].

Sensitivity and specificity are commonly used metrics to evaluate classification performance for two classes separately. Sensitivity is also known as recall or actual positive rate, while specificity is the true negative rate or negative class accuracy. However, sensitivity and specificity are trade-offs and cannot be used alone to evaluate classifier performance. Therefore, this study uses G-Mean, the geometric mean of sensitivity and specificity, to combine the classifier's ability to identify both positive and negative classes. The G-Mean is calculated using Equation (3) to evaluate the classification results.

$$G_{mean} = \text{Sensitivity} \times \text{Specificity} \qquad (3)$$

We also observed the effective detection ability for only one class by adopting another pair of metrics, precision and recall. F-Measure is used to integrate precision and recall into a single metric. F-Measure could be calculated by using the following Equation 4.

$$F_{measure} = \frac{2 \times \text{precission} \times \text{recall}}{\text{precission} + \text{recall}} \qquad (4)$$

3. Results and Discussion

This section will explain the results generated from this research, which are divided into four topics: how to determine the best feature subset for SNP identification, the influence of feature selection in SNP identification, the fact that feature selection shifts the sample position, and the impact of combining feature selection with undersampling.

3.1. *Determining the best Feature Selection in SNP Identification*

Referring to the application of the SNR technique, the selected features with significant SNR scores are regarded as the best features due to their superior ability to distinguish classes. The average SNR value of each feature from 20 chromosomes is used as the final SNR score, which is presented in Table 1. Among the chosen features, two specific features, namely the maximum quality of minor alleles and allele balance, have been previously recommended by Istiadi *et al.* (2014)] for the identification of SNPs. Additionally, allele balance has been recognized as an essential attribute in variant filtering for the detection of SNPs and insertion-deletions (INDELs) in studies related to rare human diseases [Pedersen *et al.* (2021]). Furthermore, the feature of error

probability has also been identified as a relevant attribute for recognizing true SNPs. This is in agreement with the findings of O'Fallon, Wooderchak-Donahue and Crockett (2013)], as it can capture conditions where sequencing or alignment errors result in a small portion of false base variation at a specific site. These selected features are considered important for the accurate identification of SNPs and can aid in the detection of true genetic variations in various genomic studies.

Feature Number	Name of feature	SNR Score
3	Maximum quality of minor allele	1.275458
5	Mean quality of minor allele	1.177911
7	Frequency of minor allele	0.737776
12	Error probability	0.618131
21	Allele balance	0.642246

Table 1. The result of feature selection using SNR

3.2. *The Influence of Feature Selection in SNP Identification*

To evaluate the impact of various feature subsets on classification accuracy, the performance of models trained using different subsets of the 24 available features was compared to that of the model trained on the complete set. The performance of each model was assessed using a set of predefined evaluation metrics. A summary of the experimental results is provided in Table 2.

Feature Number	Sensitivity	Specificity	G-Mean	F-Measure
3,5,7	0,59	0,96	0,75	0,62
3,5,7,21	0,63	0,96	0,78	0,63
3,5,7,21,12	**0,65**	**0,96**	**0,79**	**0,66**
All 24 features	0,60	0,97	0,76	0,63

Table 2. The planning and control components.

According to the results presented in Table 1, the model utilizing five features exhibited the highest sensitivity among all models, even surpassing the model trained on the full set of 24 features. Specifically, the model employing five features achieved a sensitivity of 0.65, indicating its ability to successfully identify 65% of positive SNPs in the testing data. Furthermore, this model also exhibited the highest values for G-Mean (0.79) and F-Measure (0.66). Notably, the computational time required for training a model using five features was 3.89 times shorter than that of a model utilizing all 24 features, as the process of selecting features effectively reduced the dataset dimensions used in the model. This resulted in improved computational efficiency, reduced computing and storage resources, and faster computation time.

These findings demonstrate that employing more features in classifier development does not necessarily translate to better results, as irrelevant or redundant features may weaken the training process, reduce classifier performance, and increase computational time [Jeni *et al.* (2013]. Hence, judicious selection of features is critical in building an effective classifier that can improve the ability to detect the minor class and expedite computation [Soufan *et al.* (2018)].

In terms of specificity, all models performed exceptionally well, with a score of 0.97 for the model with 24 features and 0.96 for the rest. This indicates that 97% and 96% of negative SNPs were accurately identified as false SNPs by the classifiers. However, a small percentage of negative SNPs (3% and 4%) were incorrectly classified as positive SNPs. This outcome reveals that all models tended to classify all instances into the negative class, disregarding the minority class, which is, in fact, a higher priority for identification. Such behavior demonstrates how imbalanced classes could impact machine learning performance, specifically with Support Vector Machines (SVM). Research has shown that as the training data becomes more imbalanced, the ratio between positive and negative support vectors becomes more imbalanced as well. Consequently, the neighborhoods of a test instance close to the boundary are more likely to be dominated by negative support vectors, causing the decision function to be more likely to classify as negative [Wu and Chang (2003)].

3.3. *Feature Selection Shift the Sample Position*

The use of data visualization is a useful tool for analyzing the structural relationships between variables in a dataset [Yang and Moody (1999)]. The feature selection process is closely intertwined with the determination of data visualization. An effective feature selection algorithm can produce a subset of features that demonstrate

meaningful visualization, such as class differentiation, using a smaller data dimension. Figure 3 compares the visualization obtained using 24 features with the five selected SNR features.

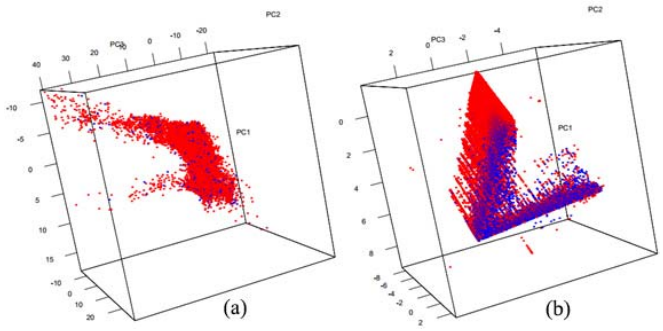

Fig. 3. Visualization of SNP candidates dataset from 16th Chromosome with (a) 24 features and (b) 5-SNR features

In Figure 4, positive SNPs are represented by light-grey dots while negative SNPs are represented by dark-grey dots. Figure 4(a) shows that the separation between positive and negative classes in the dataset is not clear, as the data from both classes appear to be blending, with no visible boundary between them. In contrast, Figure 4(b) displays a noticeable difference in the position of positive and negative class data. Feature selection provides an alternative approach to addressing imbalanced data problems because the selection of features is related to the selection of relevant data points [Shashua and Wolf (2004)]. Thus, the selection of features can shift the position and distribution of data.

3.4. *The Impact of Combining Feature Selection with Undersampling*

In this study, we also aim to investigate the impact of integrating feature selection and data balancing techniques. The data balancing approach employed in this study is 1:1 undersampling, following the recommendation of a previous study [Hasibuan *et al.* (2014)]. The dataset under analysis corresponds to chromosome 16, and it includes five features that were recommended in prior experiments. To assess the effectiveness of the proposed methodology, we evaluate its performance using various metrics, which are presented in Figure 5. These findings can provide valuable insights into the development of effective strategies for analyzing genomic data, particularly with respect to improving the accuracy of SNP identification through the utilization of feature selection and data balancing techniques.

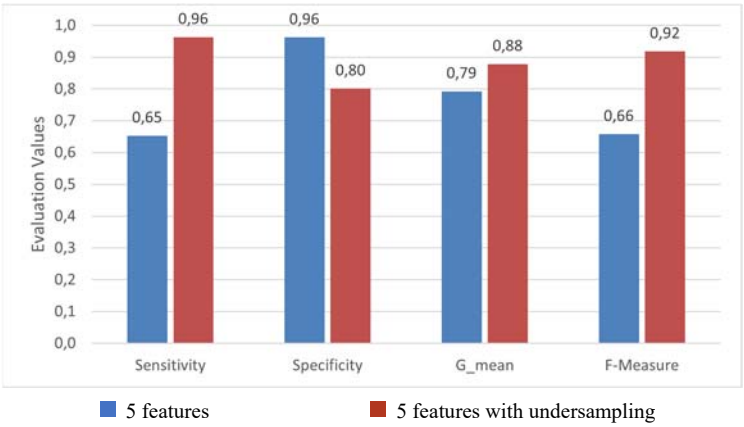

Fig. 4. Comparison of classifier performance using five features with and without undersampling.

Figure 4 presents a bar chart comparing the performance of two models in identifying SNPs. The first model utilizes feature selection alone, while the second model combines feature selection with undersampling. The implementation of undersampling resulted in an improvement in sensitivity, which increased by 47% to 0.96, indicating an increased ability to identify positive SNPs. The G-Mean value also increased by 11% to 0.88, representing an overall improvement in model performance. Furthermore, the F-measure increased by 39.5% to 0.92, implying a higher number of correctly identified positive SNPs. However, the specificity value decreased by 16% to 0.8, indicating a reduced ability to identify negative SNPs due to an increase in false positives.

The efficiency of the undersampling technique was evaluated and found to reduce the computational time required to build the model by up to 28 times compared to models without undersampling. This is due to the undersampling technique could act as a data-cleaning process, where redundant or irrelevant samples are eliminated, thereby increasing the effectiveness and efficiency of classification. Based on the comparison made, the model that combines feature selection techniques with undersampling outperforms the model utilizing feature selection alone in identifying SNPs.

Additionally, we compared classifiers using five features followed by undersampling with those using 24 features followed by undersampling, and the results are presented in Figure 5.

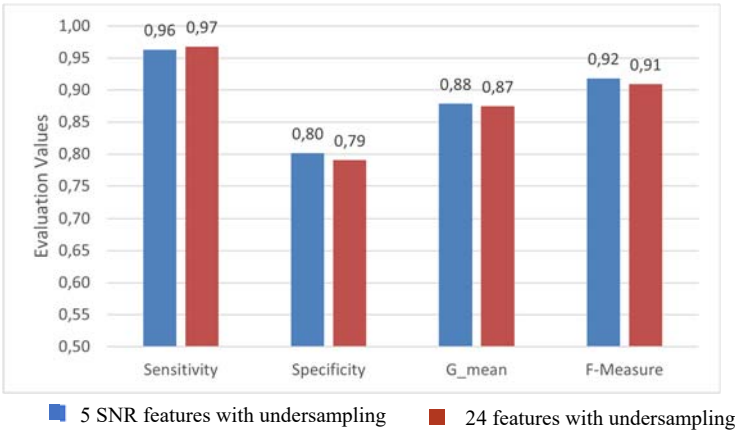

Fig. 5. Comparison of classifier performance using undersampling technique with 5 and 24 features.

The results presented in Figure 5 demonstrate that the model utilizing a reduced set of five features exhibits a slightly superior G-Mean value of 0.88, accompanied by an increase in the F-Measure value of 0.92. Of note, this model not only enhances the ability to identify positive instances, but it also requires significantly less time for computation, with a speed improvement of 4.8 times compared to the model employing 24 features. These findings highlight an additional benefit of undersampling, as this technique effectively reduces both the computation time and storage requirements, as previously reported [Yu *et al* (2013)].

4. Conclusion

In conclusion, our study demonstrates the effectiveness of combining feature selection and undersampling techniques in improving the performance of SNP identifier models. Specifically, using a combination of 5 selected features and 1:1 undersampling, we achieved the best performance with G-Mean and F-Measure values of 0.88 and 0.92, respectively, while also reducing computation time by 28 times compared to conventional methods. Our analysis also identified the five most important features for SNP identification, including the maximum quality of minor alleles, mean quality of minor alleles, minor allele frequency, error probability, and allele balance. These findings provide valuable insights for improving SNP identification methods and advancing genomic research.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets., (pp. 39-50).
[2] Cuperlovic-Culf, M., Belacel, N., & Ouellette, R. J. (2005). Determination of tumour marker genes from gene expression data. *Drug Discovery Today, 10*(6), 429-437. doi:https://doi.org/10.1016/S1359-6446(05)03393-3
[3] Ditzler, G., & Polikar, R. (2013). Incremental learning of concept drift from streaming imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 25*(10), 2283-2301. doi:10.1109/TKDE.2012.136
[4] Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling, 61*(6). doi:10.1021/acs.jcim.1c00160
[5] Golub, T. R., Slonim, †. D., Tamayo, †. P., Huard, C., Gaasenbeek, M., Mesirov, J. P., . . . Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. Retrieved from www.genome.wi.mit.edu/MPR
[6] Gunavathi, C., & Premalatha, K. (2014). A Comparative Analysis of Swarm Intelligence Techniques for Feature Selection in Cancer Classification. *Scientific World Journal, 2014*. doi:10.1155/2014/693831

[7] Gupta, P. K., Roy, J. K., & Prasad, M. (2001). Single Nucleotide Polymorphisms : A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science, 80*.
[8] Guyon, I., & De, A. M. (2003). An Introduction to Variable and Feature Selection André Elisseeff. *An Introduction to Variable and Feature Selection André Elisseeff, 3*, 1157-1182.
[9] Han, R., Wang, S., & Gao, X. (2020, March). Novel algorithms for efficient subsequence searching and mapping in nanopore raw signals towards targeted sequencing. *Bioinformatics, 36*(5), 1333-1343. doi:10.1093/bioinformatics/btz742
[10] Hasibuan, L. S., Kusuma, W. A., & Suwamo, W. B. (2014). 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*(June), 375-379. doi:10.1109/ICACSIS.2014.7065854
[11] Haury, A. C., Gestraud, P., & Vert, J. P. (2011, December). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE, 6*(12). doi:10.1371/journal.pone.0028210
[12] Huang, C.-J., & Liao, W.-C. (2003). A comparative study of feature selection methods for probabilistic neural networks in cancer classification., (pp. 451-458).
[13] Istiadi, M. A., Kusuma, W. A., & Tasma, I. M. (2014). Application of decision tree classifier for single nucleotide polymorphism discovery from next-generation sequencing data. doi:10.1109/ICACSIS.2014.7065832
[14] Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics, 9*(2), 166-177.
[15] Jeni, L. A., Cohn, J. F., & Torre, F. D. (2013). Facing imbalanced data–recommendations for the use of performance metrics., (pp. 245-251).
[16] Kirov, K., Krachunov, M., Kulev, O., Nisheva, M., Minkov, G., & Vassilev, D. (2016). Reducing false negatives for errors in SNP detection using a machine learning approach. *Comptes Rendus de L'Academie Bulgare des Sciences, 69*(7).
[17] Ko, G. H., Kim, P. G., Yoon, J., Han, G., Park, S. J., Song, W., & Lee, B. (2018). Closha: bioinformatics workflow system for the analysis of massive sequencing data. *BMC bioinformatics, 19*. doi:10.1186/s12859-018-2019-3
[18] Korani, W., Clevenger, J. P., Chu, Y., & Ozias-Akins, P. (2019, March). Machine Learning as an Effective Method for Identifying True Single Nucleotide Polymorphisms in Polyploid Plants. *The Plant Genome, 12*(1), 180023. doi:10.3835/plantgenome2018.05.0023
[19] Koziarski, M. (2021). Potential Anchoring for imbalanced data classification. *Pattern Recognition, 120*. doi:10.1016/j.patcog.2021.108114
[20] Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern recognition, 33*(1), 25-41.
[21] Kwok, P.-Y., & Chen, X. (2003). Detection of Single Nucleotide Polymorphisms 43 Detection of Single Nucleotide Polymorphisms. *Detection of Single Nucleotide Polymorphisms 43 Detection of Single Nucleotide Polymorphisms, 5*, 43-60.
[22] Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., . . . Zhang, G. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics, 42*(12), 1053-1059. doi:10.1038/ng.715
[23] Lema, M. (2018). Marker assisted selection in comparison to conventional plant breeding. *Agric. Res. Technol, 14*, 555914.
[24] Li, W., Richter, R. A., Jung, Y., Zhu, Q., & Li, R. W. (2016). Web-based bioinformatics workflows for end-to-end RNA-seq data computation and analysis in agricultural animal species. *BMC Genomics, 17*(1). doi:10.1186/s12864-016-3118-z
[25] Liao, P.-Y., & Lee, K. H. (2010). From SNPs to functional polymorphism: The insight into biotechnology applications. *Biochemical Engineering Journal, 49*(2), 149-158. doi:https://doi.org/10.1016/j.bej.2009.12.021
[26] Magana-Mora, A., & Bajic, V. B. (2017). OmniGA: Optimized Omnivariate Decision Trees for Generalizable Classification Models. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-04281-9
[27] Matukumalli, L. K., Grefenstette, J. J., Hyten, D. L., Choi, I. Y., Cregan, P. B., & Tassell, C. P. (2006). Application of machine learning in SNP discovery. *BMC Bioinformatics*. doi:10.1186/1471-2105-7-4
[28] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. C. (2014). e1071: Misc functions of the Department of Statistics (e1071), TU Wien. *R package version, 1*(3).
[29] Mishra, D., & Sahu, B. (2011). A signal-to-noise classification model for identification of differentially expressed genes from gene expression data., *2*, pp. 204-208. doi:10.1109/ICECTECH.2011.5941685
[30] Mishra, D., & Sahu, B. (2011). Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach. *International Journal of Scientific & Engineering Research, 2*(4). Retrieved from http://www.ijser.org
[31] Nurhasanah, R., Hasibuan, L. S., & Kusuma, W. A. (2020, July). Feature Selection Approach for Solving Imbalanced Data Problem in Single Nucleotide Polymorphism Discovery. *1566*. Institute of Physics Publishing. doi:10.1088/1742-6596/1566/1/012035
[32] O'Fallon, B. D., Wooderchak-Donahue, W., & Crockett, D. K. (2013). A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics*. doi:10.1093/bioinformatics/btt172
[33] Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tristani-Firouzi, M., . . . Quinlan, A. R. (2021). Effective variant filtering and expected candidate variant yield in studies of rare human disease. *npj Genomic Medicine, 6*(1), 60. doi:10.1038/s41525-021-00227-3
[34] Sahu, B., & Mishra, D. (2011). A novel approach for selecting informative genes from gene expression data using Signal-to-Noise Ratio and t-statistics., (pp. 5-10). doi:10.1109/ICCCT.2011.6075207
[35] Shashua, A., & Wolf, L. (2004). Kernel feature selection with side data using a spectral approach., (pp. 39-53).
[36] Soufan, O., Ba-Alawi, W., Afeef, M., Essack, M., Rodionov, V., Kalnis, P., & Bajic, V. B. (2015). Mining Chemical Activity Status from High-Throughput Screening Assays. *PLoS ONE, 10*(12). doi:10.1371/journal.pone.0144426
[37] Soufan, O., Ba-Alawi, W., Magana-Mora, A., Essack, M., & Bajic, V. B. (2018). DPubChem: A web tool for QSAR modeling and high-throughput virtual screening. *Scientific Reports, 8*(1). doi:10.1038/s41598-018-27495-x
[38] Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., . . . Scholten, T. (2020). Synthetic resampling strategies and machine learning for digital soil mapping in Iran. *European Journal of Soil Science, 71*(3). doi:10.1111/ejss.12893
[39] Tang, Y., Zhang, Y. Q., & Chawla, N. V. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39*(1), 281-288. doi:10.1109/TSMCB.2008.2002909
[40] Verikas, A., & Bacauskiene, M. (2002). Feature selection with neural networks. *Pattern Recognition Letters, 23*(11), 1323-1335. doi:https://doi.org/10.1016/S0167-8655(02)00081-8
[41] Wang, K. F., An, J., Wei, Z., Cui, C., Ma, X. H., Ma, C., & Bao, H. Q. (2022). Deep Learning-Based Imbalanced Classification With Fuzzy Support Vector Machine. *Frontiers in Bioengineering and Biotechnology, 9*. doi:10.3389/fbioe.2021.802712
[42] Wani, S. H., Choudhary, M., Kumar, P., Akram, N. A., Surekha, C., Ahmad, P., & Gosal, S. S. (2018). *Marker-Assisted Breeding for Abiotic Stress Tolerance in Crop Plants.* Springer International Publishing. doi:10.1007/978-3-319-94746-4_1
[43] Wasikowski, M., & Chen, X.-w. (2010). Combating the Small Sample Class Imbalance Problem Using Feature Selection. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1388-1400. doi:10.1109/TKDE.2009.187

[44] Wu, G., & Chang, E. Y. (2003). Class-boundary alignment for imbalanced dataset learning., (pp. 49-56).
[45] Yang, G., Chen, S., Chen, L., Sun, K., Huang, C., Zhou, D., . . . Guo, T. (2019, December). Development of a core SNP arrays based on the KASP method for molecular breeding of rice. *Rice, 12*(1). doi:10.1186/s12284-019-0272-3
[46] Yang, H., & Moody, J. (1999). Data visualization and feature selection: New algorithms for nongaussian data. *Advances in neural information processing systems, 12*.
[47] Yu, D. J., Hu, J., Tang, Z. M., Shen, H. B., Yang, J., & Yang, J. Y. (2013). Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. *Neurocomputing, 104*. doi:10.1016/j.neucom.2012.10.012
[48] Yu, G., Yang, Y., Yan, Y., Guo, M., Zhang, X., & Wang, J. (2022). DeepIDA: Predicting Isoform-Disease Associations by Data Fusion and Deep Neural Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19*(4). doi:10.1109/TCBB.2021.3058801
[49] Zhang, C., Soda, P., Bi, J., Fan, G., Almpanidis, G., García, S., & Ding, W. (2022). An empirical study on the joint impact of feature selection and data resampling on imbalance classification. *Applied Intelligence*. doi:10.1007/s10489-022-03772-1

Authors Profile

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, and object detection.

Agus Buono is a Professor of computer science in IPB University, Indonesia. He received his Master of Statistics from the Department of Statistics - Bogor Agricultural University, and both Master and Doctorate in Computer Science from the Faculty of Computer Science - University of Indonesia. His research mainly focuses on computational intelligence modeling applied to various fields, such as signal processing, climate modeling, modeling climate change impacts on various sectors (agriculture, health, water resources, engineering, management).

Wisnu Ananta Kusuma is an Associate Professor at the Department of Computer Science, and a researcher at Tropical Biopharmaca Research Center, IPB University, Indonesia. He is also an executive committee of the Asia Pacific Bioinformatics Network and president of the Indonesian Society of Bioinformatics and Biodiversity. His research interests include bioinformatics, machine learning, network pharmacology, and high-performance computing.