SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.03 NO. 2 2024
DOI: 10.51984/SUCP.V3I2.3357

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

Addressing Class Imbalance for Breast Cancer Prediction in Southern Libya: A Comparative Study of Sampling Techniques

*Asma Agaal[1], Mansour Essgaer[2], Amal Amarrf[2]

[1] Artificial Intelligence Department, Faculty of Technical Sciences, Sebha University, Libya,
[2] Artificial Intelligence Department, Faculty of Information Technology, Sebha University, Libya

Keywords:
Sebha Oncology Center
Libya
Random Forest
breast cancer
SMOTE
Class imbalance
majority class
sampling techniques
stratified cross-validation
Grid Search

ABSTRACT

Class imbalance refers to a scenario where the quantity of data in the minority class is significantly lower than that in the majority class, resulting in challenges in classification. To address this issue, this study tackles the challenge of class imbalance in breast cancer prediction using a dataset from the Sabha Center for Oncology Treatment in southern Libya. The research investigates the impact of eight different sampling techniques, including SMOTE, Adasyn, and NearMiss, when combined with Random Forest classification. The findings reveal that integrating SMOTE with Random Forest significantly outperforms other model configurations, resulting in a 21% increase in accuracy for predicting malignant samples and reaching a peak recall of 96%. This study demonstrates the importance of addressing class imbalances in medical datasets to improve the effectiveness of breast cancer prediction models.

معالجة اختلال التوازن الطبقي للتنبؤ بسرطان الثدي في جنوب ليبيا: دراسة مقارنة لتقنيات أخذ العينات

*اسمه اعجال[1] و منصور الصغير[2] و امال معيوف[2]

[1]قسم الذكاء الاصطناعي، كلية العلوم التقنية، جامعة سبها، ليبيا
[2]قسم الذكاء الاصطناعي، كلية تقنية المعلومات ، جامعة سبها ، ليبيا

الكلمات المفتاحية:
اختلال التوازن الطبقي
البحث الشبكي
التحقق الطبقي المتبادل
الغابة العشوائية
فئة الأغلبية
ليبيا
مركز أورام سبها
SMOTE
سرطان الثدي
تقنيات أخذ العينات

الملخص

يشير اختلال التوازن الطبقي إلى سيناريو تكون فيه كمية البيانات في فئة الأقلية أقل بكثير من تلك الموجودة في فئة الأغلبية، مما يؤدي إلى تحديات في التصنيف. ولمعالجة هذه المشكلة، تتناول هذه الدراسة التحدي المتمثل في عدم التوازن الطبقي في التنبؤ بسرطان الثدي باستخدام مجموعة بيانات من مركز سبها لعلاج الأورام في جنوب ليبيا. يبحث البحث في تأثير ثمانية تقنيات مختلفة لأخذ العينات، بما في ذلك SMOTE وAdasyn وNearMiss، عند دمجها مع تصنيف Random Forest. تكشف النتائج أن دمج SMOTE مع Random Forest يتفوق بشكل كبير على تكوينات النماذج الأخرى، مما يؤدي إلى زيادة بنسبة 21% في الدقة للتنبؤ بالعينات الخبيثة والوصول إلى ذروة الاستدعاء بنسبة 96%. توضح هذه الدراسة أهمية معالجة الاختلالات الطبقية في مجموعات البيانات الطبية لتحسين فعالية نماذج التنبؤ بسرطان الثدي.

1. Introduction

Cancer poses a growing global health threat, projected to become a leading cause of death in the near future [1]. Early detection and diagnosis are crucial for successful treatment [2], but analyzing healthcare data presents a challenge due to pervasive class imbalances [3]. This disparity, where data points for a particular condition (e.g., malignancy) are significantly outnumbered by those representing the absence of that condition (e.g., benign), hinders the development of effective prediction models.

To address this challenge, researchers have developed various techniques, primarily categorized as algorithm-level, data-level [4], or cost-sensitive approaches [5, 6]. This paper focuses on a data-level learning method that rebalances datasets using sampling techniques.

*Corresponding author:
E-mail addresses: asma.agaal@sebhau.edu.ly, (M. Essgaer) man.essgaer@sebhau.edu.ly, (A. Amarrf) ama.amarrf@sebhau.edu.ly
Article History : Received 19 June 2024 - Received in revised form 29 September 2024 - Accepted 06 October 2024

This approach, combined with stratified cross-validation and hyper parameter tuning [7], aims to enhance the performance of predictive models.

The study investigates the effectiveness of eight distinct sampling techniques, including SMOTE [8], Adasyn [9], NearMiss [10], EditedNearestNeighbours [11], Random Undersampling, SMOTETomek [12], Random Oversampling, and tomekU sampling [13], in conjunction with the Random Forest algorithm. The research utilizes a dataset of routine blood analyses from breast cancer patients at an oncology center in southern Libya.

Given the critical nature of accurately identifying malignant cases, recall is used as the primary performance metric. While precision and overall accuracy are also important, a high recall minimizes the risk of false negatives, which is crucial for early disease detection and successful treatment outcomes.

The subsequent sections of this paper will provide an in-depth review of the literature on addressing imbalanced data (Section 2), a detailed description of our methodology and the techniques employed (Section 3), an extensive presentation of our results and discussion (Section 4), and, finally, our conclusions (Section 5).

2. Literature Reviews

The challenge of class imbalance in machine learning has been extensively studied, with researchers exploring various solutions to improve classification performance in datasets where one class significantly outweighs the others.

Comparative Studies and Classification of Solutions:

Studies [14, 15] contribute to this field. [14] provides a comprehensive empirical comparison of 85 minority sampling techniques across 104 imbalanced datasets, while [15] offers a broader classification of solutions, categorizing them into preprocessing, cost-sensitive learning, and reinforcement techniques. [15] also emphasizes the importance of considering intrinsic data properties such as small elements, sparse training data, overlapping classes, noisy data, boundary instances, and dataset transformations between training and testing distributions.

Challenges and Considerations:

[16] highlights the challenge of unknown true misclassification costs in the learning phase. [17] addresses this by introducing a cost-sensitive extension of the least squares (LMS) algorithm, assigning varying weights to errors from different samples. This approach is compared to traditional sampling techniques like under-sampling, over-sampling, and SMOTE.

Validation and Performance Metrics:

[18] underscores the influence of validation strategies on classifier evaluation and the need for alternative performance metrics to balance datasets. [19] addresses the limitations of standard learning algorithms, which often assume balanced class distributions, and proposes an algorithm to enhance learning in non-relative imbalanced datasets.

Medical Data and Feature Selection:

[20] focuses on the challenges of imbalanced medical data in the context of brain tumor diagnosis, utilizing optimization-based feature selection with ensemble classification. [21] investigates the optimal order of applying feature selection and oversampling techniques, finding that combining them outperforms individual use, with Information Gain (IG) followed by SMOTE yielding the best results.

Oversampling Techniques and Algorithm Comparison:

[22] explores various methods for addressing class imbalance at different levels: data level, algorithm level, and hybrid approaches. It compares different oversampling techniques, including SMOTE, ADASYN, Borderline-SMOTE, and SMOTE. Finally, [23] examines the impact of sampling techniques on classification performance, comparing ADASYN, SMOTE, and SMOTE-ENN in conjunction with different classification algorithms such as AdaBoost, K-Nearest Neighbor (K-NN), and Random Forest. The results show that combining ADASYN with Random Forest offers a 5% to 10% improvement in classification performance.

These studies provide a foundation for the current research, which delves further into the performance of various sampling techniques and their impact on learning models.

3. Material and method

This study employs a methodological framework to address class imbalance in machine learning, specifically within the context of breast cancer prediction. The framework involves the following steps:

a. Data collection and description

The study utilized a dataset of 1,800 breast cancer cases collected from the Cancer Treatment Center in South Libya, spanning the years 2015 to 2022. Patient records were manually digitized from hardcopy files. The dataset includes 22 features, comprised of terms extracted from routine blood reports of breast cancer patients, which serve as indicators for predicting the presence of the disease as shown in Table1.

TABLE 1: dataset features

Feature	Feature description
Sex	The patient's gender
Age	The patient's age
Address	The patient's address
FBS	Blood glucose
Urea	Kidney function test urea
Creatinine	Kidney function test creatinine
ALB	Albumin
T.Ca	Total calcium in the blood
GPT	Liver functions ' gpt'
GOT	Liver functions ' got '
ALP	Alkaline Phosphate
HGB	Hemoglobin
PLT	Blood platelets
ESR	Deposition of blood
LDH	Lactate Dehydrogenizes
Na+	Sodium
K	Potassium
CL-	Chloride acid
CA-15.3	Cancer antigen
CEA	Carcinoma embryonic antigen
WBC	White blood cells
RBC	Red blood cells
CLASS	Benign=0 or Malignant=1

b. Data Preprocessing

Data preprocessing was performed on the collected dataset to prepare it for analysis. This involved:

i. *Data Cleaning:* Removing noisy or irrelevant entries, identifying outliers using the Interquartile Range (IQR) method [24], and substituting missing values with their respective averages [25].

ii. *Scaling the Data:* Applying the Robust Scaler technique to ensure features are evenly distributed and on a comparable scale, improving the performance of machine learning algorithms. [26, 27].

iii. *Stratified Cross-Validation:* Employing stratified subsampling to maintain class frequency balance in each fold of cross-validation, preventing biased model evaluation due to class imbalance in the original dataset [28].

c. Resampling methods

Resampling is a widely used technique for addressing class imbalance in machine learning. Its primary aim is to achieve a balanced distribution of samples, often aiming for a 50:50 split between the minority and majority classes [29]. A key advantage of resampling is its compatibility with standard learning algorithms, as it can be implemented without requiring changes to the algorithms themselves [30]. This flexibility makes resampling a practical approach [31]. This paper focuses on using oversampling, under sampling, and data pre-processing techniques to achieve a balanced distribution. Several algorithms are employed within these resampling methods, including: *Oversampling* [32], *Under-sampling* [33] and hybrid approach [34].

Oversampling technique

Oversampling aims to rectify imbalance by introducing synthetic samples to the minority class. This can involve replicating existing samples or generating new ones. Replication can be random or target boundary samples, encouraging the classifier to allocate these regions to the minority class. Critics argue that oversampling simply rebalances distribution without adding novel insights. To address this, techniques generate new synthetic samples within plausible regions [35] Some prominent oversampling methods include:

i. Random Over Sampler: Balances class distribution by randomly duplicating instances from the minority class until it

matches the number of instances in the majority class. This involves selecting a number of minority class instances equal to the size of the majority class and adding these copies to the original dataset [36].

ii. Adaptive Synthetic Sampling (ADASYN): It is a more sophisticated oversampling technique that considers the density distribution of data points. It balances the dataset by combining random oversampling with the generation of synthetic samples, focusing on areas where the minority class is less dense[37].

iii. Synthetic Minority Over-Sampling Technique (SMOTE): it generates synthetic samples to augment the minority class by interpolating between existing minority class samples and their nearest neighbours. [32] It creates new samples by taking a linear combination of a minority class sample (X) and its nearest neighbour (Y), using a randomly generated value between 0 and 1 to determine the position of the new sample along the line connecting X and Y. A value of 0 produces a sample identical to X, while a value of 1 produces a sample identical to Y [38]. The basic Eq.1 for generating a synthetic sample S is:

$$S = X + (Y - X) * random\_number \quad (1)$$

iv. SMOTE-TL (SMOTE for the Topology-Preserving Synthetic Sample Generation): It SMOTE-TL builds on SMOTE, focusing on preserving minority class topology during synthetic sample creation. It identifies borderline & noisy samples based on density, sets a "topological level" & distance threshold, then generates new samples using a seed & its nearest neighbours to maintain similarity. Cleaning steps remove disruptive noise & borderline samples [39].

Under-sampling techniques

It balances datasets by removing redundant samples from the majority class to achieve a predetermined ratio. This can be done randomly or by targeting borderline samples, reducing majority class allocation & improving minority class classification. However, information loss is a potential drawback [40]. Prominent under sampling techniques include:

i. **Random Under Sampler**: This method reduces the majority class by randomly selecting and removing instances from the majority class to achieve a balanced class distribution [41].

ii. **Tomek-Links**: It identify pairs of samples close to each other but belonging to different classes. Using a distance metric like Euclidean distance. a Tomek-Link forms when two samples from different classes are each other's nearest neighbours. These pairs represent samples close to the decision boundary between classes.[12].

iii. **NearMiss:** It reduces imbalance by retaining only majority class samples closest to minority class samples. It calculates distances between each majority class sample & its nearest minority class sample, then selects majority class samples with the shortest distances, adjusting the number for balance. The final dataset includes these selected majority class samples & the entire minority class, creating a more balanced dataset for machine learning models [42].

iv. **EditedNearestNeighbours**: It identifies samples with nearest neighbors of different classes, focusing on majority class samples with a significant number of different-class neighbors, which are considered potentially misclassified. These samples are pruned, removing potentially noisy or incorrectly labelled data. The remaining dataset is considered cleaner, with fewer misclassified majority class samples [43].

d. Random Forest algorithm (RF)

Random Forest is an ensemble learning method that combines multiple decision trees, each trained on a random subset of the data and features. [39] Each tree recursively splits the data based on chosen features using criteria like Gini impurity or entropy. The final prediction is an ensemble average of the individual tree predictions, often using a majority vote for classification or averaging for regression [44]. This can be represented mathematically as shown in Eq.2:

$$\boldsymbol{F(x) = (1/N) * \Sigma[\, H(x)\,]} \quad (2)$$

where H(x) is an individual tree's prediction and N is the number of trees.

e. Grid Search Cross-Validation

Grid Search CV optimizes machine learning models by systematically searching through a range of hyper parameter values to find the best combination. It evaluates each combination using cross-validation and selects the set that performs best on validation data, which is then used to train the final model [45].

f. Evaluation Metrics

Model performance is evaluated using metrics like the confusion matrix, precision, recall, accuracy, F1 score, and AUC-ROC, which assess aspects like false alarm rate, ability to capture relevant instances, overall correctness, balance between precision and recall, and class distinction capability[46].

4. Results and discussions

This section presents the results of the proposed framework for breast cancer (BC) prediction using a dataset from the SOC. The study utilizes the Python environment for its machine learning capabilities. After data pre-processing, a five-fold stratified cross-validation approach was used to evaluate different methods. The experiments involved:

a. Baseline Random Forest: The initial performance of the Random Forest algorithm was assessed.

b. Random Forest Enhancements: The Random Forest algorithm was enhanced using several methods, including:
- Stratified K-Fold cross-validation.
- Various data balancing techniques.
- Grid Search CV for hyper parameter tuning.

c. Optimal Data Balancing Technique: The effectiveness of different data balancing techniques was compared to determine the optimal method.

d. Performance Comparison: The performance of the enhanced Random Forest model was compared to the baseline to quantify the improvement achieved.

Insight into the data

The dataset used in the study exhibited several challenges, including a high number of missing values as shown in Fig 1, outliers as shown in Fig 2, and a significant class imbalance as shown in Fig 3, with a majority of benign samples (1351) and a smaller number of malignant samples (449).

Fig. 1. Missing values for each attribute in the data set

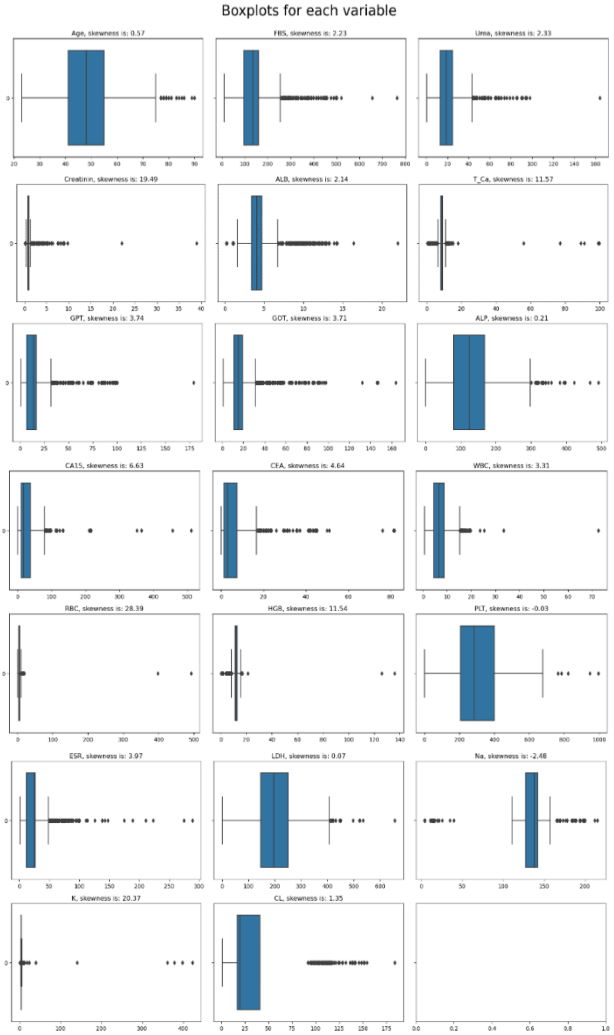

Fig. 2. Outliers for each attribute in the data set

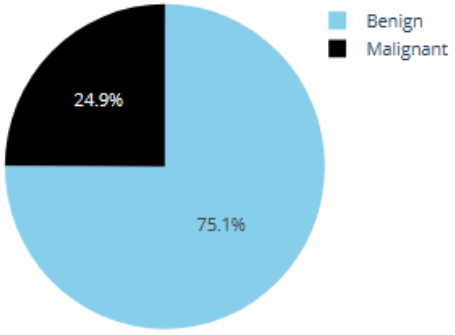

Fig. 3. Distribution of target attribute in the study data

To address the data challenges, preprocessing steps were applied. Missing values were imputed using the arithmetic mean, outliers were handled using the Interquartile Range (IQR) method, and data was standardized using Robust Scaler (RS).

Initial Random Forest Performance

The initial Random Forest model, without any optimization, achieved precision of 80%, accuracy of 75%, recall of 70%, and an F1 score of 87%. However, these results may be limited by the class imbalance in the dataset, suggesting potential for improvement, which will be detailed in the subsequent sections.

1) ***Boosting the Random Forest Algorithm's Performance***

The model's performance can be enhanced by improving both the data quality and fine-tuning the algorithm's parameters. As shown below:

Applying Stratified K Fold

Stratified K-Fold cross-validation divides the data into five parts, using one for testing and the rest for training, to improve model evaluation and prevent overfitting.

Applying Grid Search

The performance of the Random Forest algorithm exhibited a substantial improvement, notably without any resampling, following a meticulous tuning of hyperparameters through a grid search approach. The enhancement was significant, with key metrics reflecting a remarkabl increase of between 5% and 10%. The refined model achieved exceptional results with a recall of 0.75, precision of 0.89, an F1 score of 0.85, and an impressive accuracy rate of 0.87. This adjustment underscores the critical role that hyperparameter tuning can play in optimizing the model's predictive abilities.

Applying resampling data methods

To improve the Random Forest algorithm's effectiveness, the study used various resampling techniques to achieve a balanced representation of benign and malignant samples. This involved applying oversampling and under sampling methods to the training data, along with stratified cross-validation and grid search hyper parameter tuning, to ensure a well-balanced and finely tuned model for optimal prediction.

- **Investigation Oversampling techniques**

For the sake of comparison, four Oversampling techniques were employed, which include RandomOverSampler, ADASYN, SMOTE, and SMOTE-TL. The alteration in the distribution of positive and negative classes in the training data can be observed in Table 2 before and after the application of these resampling methods.

TABLE 2. NUMBER OF CATEGORIES FOR EACH OVERSAMPLING TECHNIQUES

Oversampling Methods	Number of Benign	Number of Malignant
RandomOverSampler	1089	1089
ADASYN	1089	1069
SMOTE	1089	1089
SMOTE-TL	943	943

After employing various oversampling techniques to tackle class imbalance, we observed distinct adjustments in the sampling distribution. The RandomOverSampler method duplicated instances from the minority class, originally consisting of 431 samples, to match the majority class with 1,351 samples. Consequently, the majority class decreased to 1,089 samples, while the minority class increased to the same number, resulting in equal representation. Similarly, SMOTE generated synthetic instances via interpolation, achieving the same balanced representation as RandomOverSampler. On the other hand, ADASYN adapted the oversampling degree for each sample, focusing on generating synthetic instances for challenging minority cases. This resulted in an increase in the minority class to 1,069 samples and a reduction of the majority class by 262 samples to 1,089. Meanwhile, SMOTE-TL concentrated on producing synthetic samples near decision boundaries, leading to 943 samples in both the majority and minority classes. Fig 4 visually represents these adjustments, highlighting how each technique affected the distribution of categories

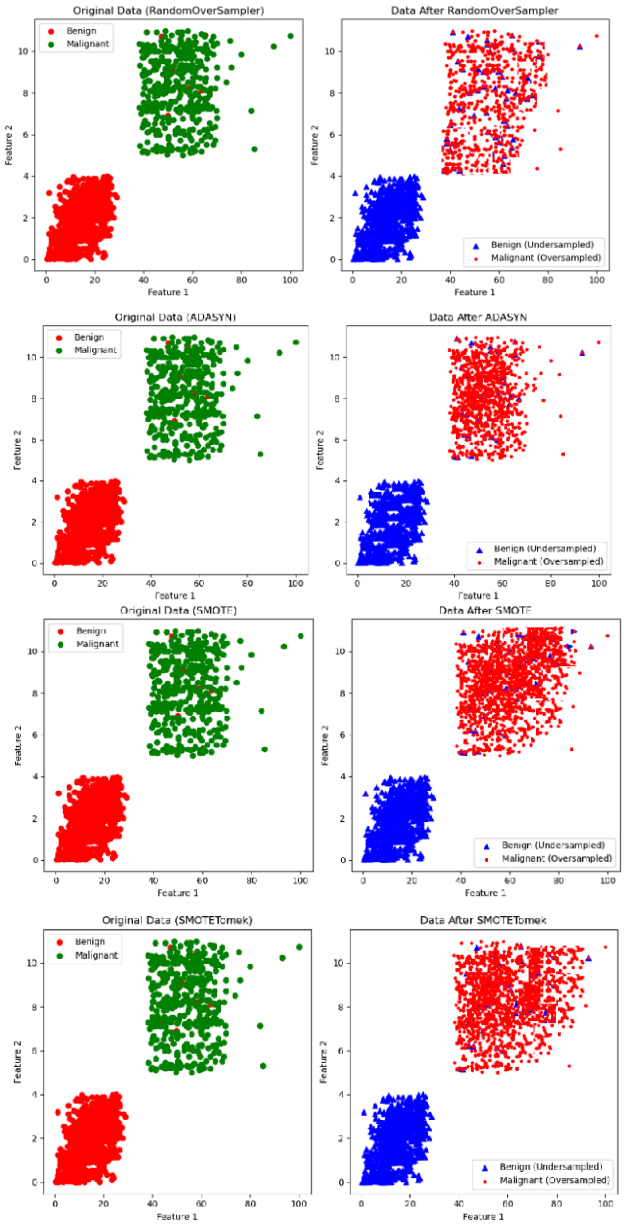

Fig. 4. Use oversampling to balance class data

- **Investigation Under Sampler techniques**

Additionally, four Under sampling techniques were implemented, namely Random Under Sampler, TomekLinks, NearMiss, and EditedNearestNeighbours. The shift in the distribution of positive and negative class instances within the training dataset is evident in Table 3, both prior to and following the utilization of these resampling approaches.

TABLE 3. NUMBER OF CATEGORIES FOR EACH UNDER SAMPLER TECHNIQUES

Under sampling Methods	Number of Benign	Number of Malignant
RandomUnderSampler	431	431
TomekLinks	900	606
NearMiss	431	314
EditedNearestNeighbours	1049	431

Through the application of various under sampling techniques, significant adjustments in the distribution of samples have been made to combat class imbalance. Let's delve into the specific impact of each method: Firstly, the Random Under Sampler approach reduces the majority class by randomly eliminating instances. In this case, the majority class shrinks from 1,089 to 431, while the minority class remains unaltered. TomekLinks, on the other hand, identifies and removes pairs of nearest neighbors with different classes. This targeted removal strategy reduces the majority class from 1,075 to 900 and trims the minority class to 606. NearMiss, another under sampling method, selectively removes samples from the majority class that are in close proximity to the minority class. This process resulted in a reduction of the majority class to 431 and a decrease in the minority class to 341. Lastly, EditedNearestNeighbors trims the majority class by discarding instances that have a substantial number of neighbors from different classes. In this instance, the majority class is downsized to 1,049, while the minority class remains at 431. Fig 5 visually illustrates these modifications, offering a clear depiction of how each technique influenced the distribution of categories, while keeping the minority class largely intact.

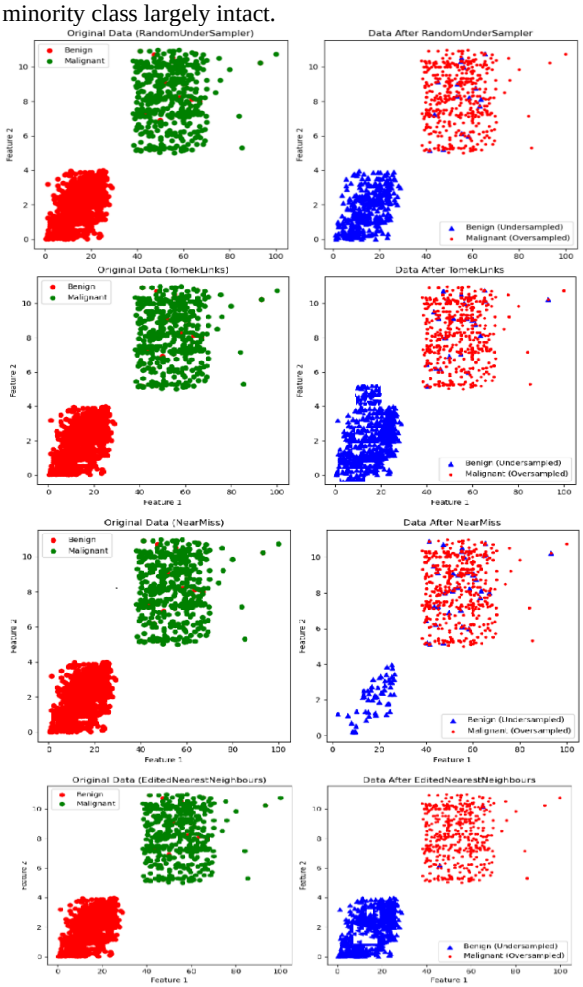

Fig. 5. Use undersampling to balance class data

2) ***Identify The Optimal Resampling data methods***

The study meticulously outlined the specific hyperparameter values for the Random Forest algorithm under each resampling strategy, revealing the unique configurations required to optimize performance with different data balancing techniques. This analysis demonstrates how the algorithm is fine-tuned to accommodate the characteristics of each resampling method, ensuring a well-balanced and optimally performing model, meticulously outlined in Table 4.

TABLE 4. HYPER PARAMETER VALUES TO THE RANDOM FOREST ALGORITHM FOR EACH SAMPLING TECHNIQUE

data resampling techniques	Max depth	N estimators	Random state
RandomOverSampler	6	200	10
ADASN	6	100	30
SMOTE	6	200	10
SMOTE-TL	12	200	10
RandomUnderSampler	6	50	30
TomekLinks	12	200	30/
NearMiss	10	200	10
EditedNearestNeighbours	10	200	20

Subsequently, the primary objective was to assess the model's accuracy in predicting samples, and this was carried out by leveraging the confusion matrix. The results of these evaluations vividly showcased the improved performance of the Random Forest model. These enhanced results are exhaustively presented in Table 5.

TABLE 5. CONFUSION MATRIX RESULTS OF THE RANDOM FOREST ALGORITHM WITH EACH SAMPLING TECHNIQUE

	TP	TN	Sum	FP	FN	Sum
RandomOverSampler	127	398	525	7	8	15
ADASN	129	397	528	8	6	14
SMOTE	130	398	528	7	5	12
SMOTE-TL	128	399	527	6	7	13
RandomUnderSampler	129	392	521	13	6	19
TomekLinks	124	400	524	5	11	16
NearMiss	129	393	522	12	6	18
EditedNearestNeighbours	127	400	527	5	8	13

Table 6 shows the Random Forest classifier's performance with various data sampling techniques, highlighting the number of correctly classified cases, True positive (TP) and True negative (TN). The combination of SMOTE and Random Forest demonstrates significant performance improvement, achieving the highest number of TP (528) and the lowest number of False negative (FN) is (5). The combination of Random Forest with ADASYN, NearMiss, and Random Under Sampler also shows good performance, each resulting in (6) FN, with ADASYN achieving the highest TP (528), followed by NearMiss (522) and Random Under Sampler (521).

SMOTE-TL ranks third, achieving a commendable 7 FN, while TomekLinks performs last with 11 FN. These results highlight the significant impact of different sampling techniques on Random Forest performance. SMOTE and ADASYN are particularly effective in increasing TP predictions while minimizing FN, which is crucial in many classification tasks.

The performance of the Random Forest algorithm is evaluated based on Recall scores, which prioritize the ability to correctly identify minority class instances. The results, presented in Table 6, provide a ranking of the sampling techniques, allowing for a data-driven selection of the most effective strategies.

TABLE 6. THE PERFORMANCE OF THE RANDOM FOREST ALGORITHM WITH SAMPLING TECHNIQUES

Random Forest with	Recall	Precision	F1 Score	Accuracy
SMOTE Oversampling	0.962963	0.948905	0.955882	0.977778
Adasyn	0.955556	0.941606	0.948529	0.974074
NearMiss	0.955556	0.914894	0.934783	0.966667
EditedNearesttNeighbours	0.955556	0.914894	0.934783	0.966667
Random Undersampling	0.955556	0.908451	0.931408	0.964815
SMOTETomek	0.948148	0.955224	0.951673	0.975926
Random Oversampling	0.940741	0.947761	0.944238	0.972222
tomekU	0.918519	0.961240	0.939394	0.970370

Table 7 shows that SMOTE Oversampling results in the highest recall (0.96%) for positive samples, especially malignant cases. Adasyn, NearMiss, EditedNearestNeighbours, and Random Undersampling achieve a recall of 0.95%, with Adasyn showing the highest precision for benign samples (0.94%). SMOTETomek and Random Oversampling rank third with a recall of 0.94%, while tomekU achieves the lowest recall (0.91%) but the highest precision for benign samples (0.96%). All sampling algorithms achieve an accuracy of 0.97%, except for NearMiss, EditedNearestNeighbours, and Random Under sampling, which reach 0.96%.

3) ***Optimization level in random forest algorithm***

The combination of sampling techniques, cross-validation, and hyper parameter tuning significantly improved the Random Forest algorithm's performance compared to the original model or using only hyper parameter tuning with cross validation, as shown in Fig 6.

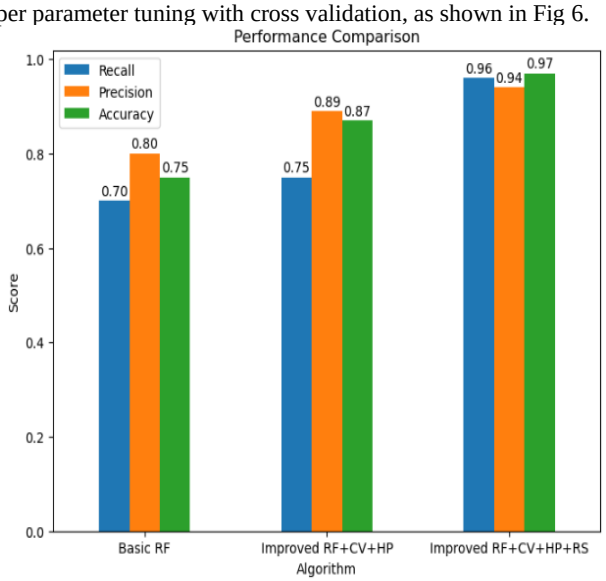

Fig. 6. Comparing the results of expert classification with the results of labeling by clustering algorithms.

The study demonstrates significant performance improvements through each stage of algorithm enhancement. While initial improvements were seen with cross-validation and hyperparameter tuning, the inclusion of sampling techniques led to even greater gains, particularly in predicting malignant samples. The combined approach resulted in a 21% increase in accuracy for malignant predictions and a 7% improvement for benign predictions. This highlights the importance of addressing class imbalance, as the algorithm's performance without sampling was misleading, primarily predicting the majority class and misclassifying the minority class.

5. Conclusion and Recommendations

This study demonstrates that incorporating sampling methods alongside cross-validation and hyperparameter tuning significantly improves classification performance, particularly for imbalanced datasets. While the algorithm without sampling may appear promising, it misclassifies the minority class due to its focus on predicting the majority. The combination of SMOTE and Random Forest proves most effective, outperforming other methods due to its ability to avoid duplicate sample values. The findings provide valuable insights for future research on handling class imbalance, including exploring larger datasets, multiclass scenarios, and integration with diverse classification algorithms.

References

[1]- Yang, F., et al., *Global trajectories of liver cancer burden from 1990 to 2019 and projection to 2035.* 2023. **136**(12): p. 1413-1421.

[2]- Jain, L., *Artificial Intelligence and Machine Learning for Healthcare.* 2023.

[3]- Jiang, Y., C. Wang, and S. Zhou. *Artificial Intelligence-based Risk Stratification, Accurate Diagnosis and Treatment Prediction in Gynecologic Oncology*. in *Seminars in Cancer Biology*. 2023. Elsevier.

[4]- Twomey, D., *Novel Algorithm-Level Approaches for Class-Imbalanced Machine Learning*. 2023, UCL (University College London).

[5]- Aguiar, G., B. Krawczyk, and A.J.M.L. Cano, *A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework.* 2023: p. 1-79.

[6]- Teslenko, D., et al., *Comparison of Dataset Oversampling Algorithms and Their Applicability to the Categorization Problem.* 2023(2 (24)): p. 161-171.

[7]- Yu, T. and H.J.a.p.a. Zhu, *Hyper-parameter optimization: A review of algorithms and applications.* 2020.
[8]- Brandt, J. and E. Lanzén, *A comparative review of SMOTE and ADASYN in imbalanced data classification.* 2021.
[9]- Qing, Z., et al., *ADASYN-LOF Algorithm for Imbalanced Tornado Samples.* 2022. **13**(4): p. 544.
[10]- Mqadi, N.M., N. Naicker, and T.J.M.P.i.E. Adeliyi, *Solving misclassification of the credit card imbalance problem using near miss.* 2021. **2021**(1): p. 7194728.
[11]- Vuttipittayamongkol, P. and E.J.I.S. Elyan, *Neighbourhood-based undersampling approach for handling imbalanced and overlapped data.* 2020. **509**: p. 47-70.
[12]- Hairani, H., A. Anggrawan, and D.J.J.I.J.o.I.V. Priyanto, *Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link.* 2023. **7**(1): p. 258-264.
[13]- Dal Pozzolo, A., et al. *Racing for unbalanced methods selection*. in *Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*. 2013. Springer.
[14]- Kovács, G.J.A.S.C., *An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets.* 2019. **83**: p. 105662.
[15]- López, V., et al., *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics.* 2013. **250**: p. 113-141.
[16]- Ishaq, A., et al., *Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques.* 2021. **9**: p. 39707-39716.
[17]- Belarouci, S., et al., *Comparative study of balancing methods: case of imbalanced medical data.* 2016. **21**(3): p. 247-263.
[18]- Raeder, T., et al., *Learning from imbalanced data: Evaluation matters.* 2012: p. 315-331.
[19]- Rendon, E., et al., *Data sampling methods to deal with the big data multi-class imbalance problem.* 2020. **10**(4): p. 1276.
[20]- Huda, S., et al., *A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis.* 2016. **4**: p. 9145-9154.
[21]- Huang, M.-W., et al., *On combining feature selection and over-sampling techniques for breast cancer prediction.* 2021. **11**(14): p. 6574.
[22]- Fotouhi, S., S. Asadi, and M.W.J.J.o.b.i. Kattan, *A comprehensive data level analysis for cancer diagnosis on imbalanced data.* 2019. **90**: p. 103089.
[23]- Kaope, C. and Y.J.M.J.M. Pristyanto, Teknik Informatika dan Rekayasa Komputer, *The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance.* 2023. **22**(2): p. 227-238.
[24]- Vinutha, H., B. Poornima, and B. Sagar. *Detection of outliers using interquartile range technique from intrusion dataset*. in *Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA*. 2018. Springer.
[25]- Little, R.J. and D.B. Rubin, *Statistical analysis with missing data*. Vol. 793. 2019: John Wiley & Sons.
[26]- Raju, V.G., et al. *Study the influence of normalization/transformation process on the accuracy of supervised classification*. in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. 2020. IEEE.
[27]- Billot, B., et al., *Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets.* 2023. **120**(9): p. e2216399120.
[28]- Mahesh, T., et al., *The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification.* 2023. **4**: p. 100247.
[29]- Yuan, Y., et al., *Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring.* 2023. **126**: p. 106911.
[30]- Stracqualursi, E., et al., *Systematic review of energy theft practices and autonomous detection through artificial intelligence methods.* 2023. **184**: p. 113544.
[31]- Kim, A. and I.J.P.o. Jung, *Optimal selection of resampling methods for imbalanced data with high complexity.* 2023. **18**(7): p. e0288540.
[32]- Wongvorachan, T., S. He, and O.J.I. Bulut, *A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining.* 2023. **14**(1): p. 54.
[33]- Singh, P.S., et al., *Enhanced classification of hyperspectral images using improvised oversampling and undersampling techniques.* 2022. **14**(1): p. 389-396.
[34]- Kou, G., et al., *Improved hybrid resampling and ensemble model for imbalance learning and credit evaluation.* 2022. **7**(4): p. 511-529.
[35]- Saalim, M.I., *Studying the perturbation-based oversampling technique for imbalanced classification problems.* 2023.
[36]- Mesquita, F., J. Maurício, and G. Marques. *Oversampling techniques for diabetes classification: A comparative study*. in *2021 International Conference on e-Health and Bioengineering (EHB)*. 2021. IEEE.
[37]- Halim, A.M., et al., *Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets.* 2023. **5**(1): p. 246− 253-246− 253.
[38]- Elreedy, D. and A.F.J.I.S. Atiya, *A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance.* 2019. **505**: p. 32-64.
[39]- Chen, B., et al., *RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise.* 2021. **553**: p. 397-428.
[40]- Tyagi, A.K. and V.K. Reddy, *Performance analysis of under-sampling and over-sampling techniques for solving class imbalance problem.* 2019.
[41]- Sarkar, S., et al. *An ensemble learning-based undersampling technique for handling class-imbalance problem*. in *Proceedings of ICETIT 2019: Emerging Trends in Information Technology*. 2020. Springer.
[42]- Tanimoto, A., et al., *Improving imbalanced classification using near-miss instances.* 2022. **201**: p. 117130.
[43]- Ludera, D.T. *Credit card fraud detection by combining synthetic minority oversampling and edited nearest neighbours*. in *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*. 2021. Springer.
[44]- Palimkar, P., R.N. Shaw, and A. Ghosh. *Machine learning technique to prognosis diabetes disease: Random forest classifier approach*. in *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021*. 2022. Springer.
[45]- Shekar, B. and G. Dagnew. *Grid search-based hyperparameter tuning and classification of microarray cancer data*. in *2019 second international conference on advanced computational and communication paradigms (ICACCP)*. 2019. IEEE.
[46]- Padilla, R., S.L. Netto, and E.A. Da Silva. *A survey on performance metrics for object-detection algorithms*. in *2020 international conference on systems, signals and image processing (IWSSIP)*. 2020. IEEE.