ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFECTIVE FILTERING EXTENSION TECHNIQUE OF SMOTE FOR CONTROLLED SYNTHETIC DATA GENERATION

SOMIYA ABOKADR[1,], AZREEN AZMAN[2], HAZLINA HAMDAN[3], NURUL AMELINA NASHARUDDIN[4]

[1,2,3,4]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400, Serdang, Malaysia

[1]Faculty of Science, Al Zintan University, Jabal al Gharbi, Libya

E-mail: [1]soma.almoktar@gmail.com, [2]azreenazman@upm.edu.my, [3]hazlina@upm.edu.my, [4]nurulamelina@upm.edu.my

Corresponding authors: SOMIYA ABOKADR, AZREEN AZMAN

ABSTRACT

Imbalanced dataset poses a significant classification challenge within the realm of machine learning and has gained growing prominence due to the need to handle real-world data that is usually imbalanced and skewed. Numerous resample techniques have been proposed in the literature to improve the performance by solving the imbalanced problem. But this situation becomes more complex when one class contains a significantly larger number of examples compared to the other classes. Under such circumstances, machine learning-based algorithms can accurately identify examples from the majority class but often struggle or likely fail to recognize instances from the minority class. However, these minority class examples often contain crucial and valuable information. In addition, generating new instances to balance the minority classes by curating some rules by the domain expert as in the standard SMOTE, may not be suitable for some instances leading to misclassification by the model. Therefore, this paper proposed a novel technique namely filter extension of the SMOTE algorithm based on optimised kernel trick on SVM to control the generation of balanced synthetic data in overlapped samples. The main objective of this paper is to predict the minority class instances accurately and robustly addressing the imbalance and overlapping issues. The proposed technique is validated by a rigorous testing framework utilizing a 10-fold cross-validation method to ensure a comprehensive evaluation of support vector machine (SVM) classifier. Several parameters, such as, AUC and G-mean metrics were used to ensure the accuracy, robustness and effectiveness of the proposed technique comparing to other traditional and machine learing-based methods. We experimented with highly imbalanced dataset from KEEL repository. The proposed approach outperformed the standard SMOTE and RC-SMOTE, proven the effectiveness of the proposed approach of filtering imbalance data in improving the generalizability and performance of machine learning classifiers.

Keywords: *SMOTE, Imbalance, Filtering, G-mean, Machine Learning*

1. INTRODUCTION

Data that has a skewed distribution is called imbalanced. More precisely, a class or classes that have fewer instances than the others. Classes with a small number of instances are usually referred to as minority classes, while those with a high number of occurrences are referred to as majority classes. There are several factors that can adversely impact the performance of classification in cases of imbalanced distribution. Key among these are issues like overlapping classes, small disjuncts, and the presence of noise instances. These challenges can complicate the classification process, as they add layers of difficulty to accurately identifying and categorising the minority class. For instance, overlapping classes can lead to ambiguity in classification boundaries, while small disjuncts and noise instances can result in misclassifications or overfitting. These issues are visually exemplified in Figure 1., providing a clearer understanding of how they manifest in imbalanced datasets.

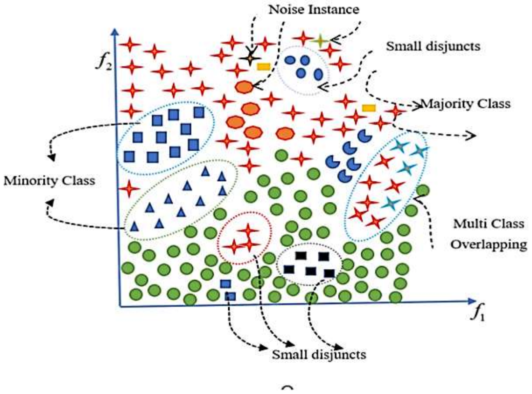

Figure 1: Imbalance Dataset Challenges

Machine learning is continually evolving to overcome inherent challenges, with imbalanced datasets being a particularly stubborn obstacle. These imbalances can heavily bias a model's performance towards the majority class, resulting in a systematic misrepresentation of minority classes. Moreover, the prevalence of class overlap muddies the waters further as classifiers struggle to distinguish between classes that are not distinctly partitioned. In light of these issues, synthetic data generation has surfaced as a critical solution, with techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) leading the charge. SMOTE and its sophisticated derivatives, such as Borderline-SMOTE and ADASYN, specifically address the class imbalance by enriching the dataset with synthetic instances strategically placed to foster a more equitable class representation. This in-depth introduction paves the way for a nuanced exploration of synthetic data generation's role in refining machine learning models, spotlighting the strategic and methodical enhancements designed to combat the intertwined challenges of imbalance and overlap within multi-class datasets.

This paper delves into the persistent challenges in machine learning related to classifying multi-class imbalanced datasets, especially those with overlapping classes. Despite considerable advancements, current methodologies often fall short of accurately distinguishing minority classes in such complex datasets. This leads to skewed predictions and a lack of reliability, particularly in critical applications where precise classification is paramount. In addition, generating new instances to balance the minority classes by curating some rules by the domain expert as in the standard SMOTE, may not be suitable for some instances, leading to misclassification by the model [1]

The paper aims to explore these challenges thoroughly, critically examine existing solutions, and introduce an innovative approach that extends the principles of synthetic data generation. The goal is to enhance the efficacy and accuracy of machine learning models in handling these intricate classification scenarios.

The contribution of this research paper is the proposal of a filter extension of the SMOTE algorithm based on an optimised kernel trick on SVM to control the generation of balanced synthetic data in overlapped samples in an imbalanced data problem. This strategy aims to bolster the efficiency and reliability of models dealing with imbalanced class classification.

2. RELATED WORK

To solve the imbalanced problem, various resample techniques have been proposed by the researchers. Resampling involves modifying the original dataset to create a more balanced class distribution, thereby improving the performance of classification algorithms. Resampling can be primarily categorized into two types: oversampling the minority class and under-sampling the majority class [2]. On the one hand, under sampling the majority class is a strategy that reduces the number of instances in the majority class [3]. Approaches such as random under sampling or the informed under sampling are applied. On the other hand, oversamplingthe minority class involves increasing the number of instances in the minority class to balance the dataset [4]. Random oversampling duplicates random samples from the minority class. While straightforward, it can lead to overfitting as it simply replicates the minority class instances. The Synthetic Sample Generation is more sophisticated techniques like SMOTE (Synthetic Minority Over-sampling Technique) and its variations (e.g., Borderline-SMOTE, ADASYN) that generate synthetic samples based on the existing minority instances. By creating new, plausible examples, these methods help the model learn a more general representation of the minority class without the overfitting risks associated with simple replication [5], as illustrated in Figure 2.

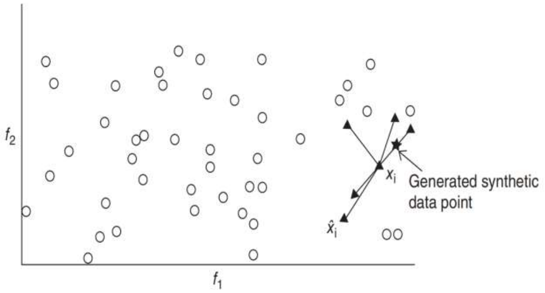

Figure 2: 2D Feature Space Using SMOTE

The synthetic minority oversampling technique (SMOTE) is a pioneering oversampling method in the study of imbalanced classification. The basic idea of SMOTE is oversampling by creating a synthetic instance in feature space formed by the instance and its K-nearest neighbours due to the ability to avoid overfitting and assist the classifier in finding decision boundaries between classes. Many extensions of SMOTE exist in recent research. In the work presented by [6], the problem of noisy and borderline examples in imbalanced classification is addressed through a re-sampling method with filtering. The authors propose SMOTE-IPF, a synthesis of the Synthetic Minority Over-sampling Technique (SMOTE) with an Iterative Partitioning Filter (IPF), designed to enhance the treatment of imbalanced data. A comprehensive experimental study is conducted, in which SMOTE-IPF is compared against other SMOTE generalizations across various synthetic and real-world datasets. The effectiveness of SMOTE-IPF in managing noisy and borderline examples is demonstrated, showing superior performance to existing SMOTE generalizations in all the different scenarios tested. The authors' analysis also sheds light on the distinctive characteristics of IPF that set it apart from other filtering approaches. Yet, the work [7], the paper titled "Dynamic ensemble selection for multi-class imbalanced datasets" is examined. This research was conducted with a focus on multi-class imbalance problems, which are significantly more complex than binary imbalanced issues. The authors developed a novel and efficacious method known as DES-MI, Dynamic Ensemble Selection for Multi-Class Imbalanced Datasets, where the competency of candidate classifiers is assessed with weighted instances in their neighbourhood. The DES-MI process involves the generation of balanced training datasets through a pre-processing routine based on random balance, followed by the selection of classifiers using a weighting mechanism to emphasize their competence, especially for classifying examples in regions with underrepresented classes. A thorough experimental study validates the effectiveness of DES-MI, showing an improvement in classification performance for multi-class imbalanced datasets, as corroborated by appropriate statistical analysis.

The study by Mehmood and Asghar in [8] introduces a modified SVM, MSVM, used as a base classifier with the AdaBoost ensemble classifier (MSVM-AdB). This approach divides the multi-class dataset into overlapping and non-overlapping regions, filtering these into critical and less critical areas based on sample contributions. The MSVM modifies the kernel mapping function of the standard SVM to map overlapped samples in a higher dimension, enhancing the classifier's capacity for learning in complex datasets. Their experimental results, utilizing 20 real datasets with varied imbalance ratios and degrees of overlap, demonstrate the superiority of MSVM-AdB over standard classifiers.

However, the evolving nature of data and the increasing sophistication of classification tasks underscore the ongoing need for innovative approaches in machine learning. As the field continues to progress, the integration of advanced techniques and methodologies will play a pivotal role in overcoming the challenges of imbalanced data classification and paving the way for more accurate, efficient, and equitable decision-making processes in various domains.

Furthermore, despite the filtering extension been one of the most widely used SMOTEs. The current studies indicate a set of rules on generated synthetic data to filter it when it is generated in a forbidden region, i.e., the majority region. However, this rule can be provided by experts: most of the nearest neighbour comes from the minority class. This rule can also be generated based on machine learning (ML). Today, many fields, such as natural language processing, have recently used the ML approach to sentiment analysis instead of traditional linguistic rules. This paper proposed filter extension of the SMOTE algorithm based on an optimised kernel trick on SVM to control the generation of balanced synthetic data in overlapped samples in an imbalanced data problem.

3. METHODOLOGY

The algorithm for the Machine Learning-Based SMOTE Filtering Extension operates as a sophisticated mechanism to enhance synthetic data generation, thereby tackling the issue of imbalanced datasets. The procedure commences with the normalisation of the dataset, utilising a specific mathematical equation to scale all features to a uniform range. The dataset is then split into distinct training and testing sets. The algorithm's core lies in the iterative balancing of the training data. It begins by selecting random instances from the minority class and identifying their nearest neighbours. Synthetic instances are generated through interpolation between these neighbours, effectively augmenting the minority class. Each new synthetic instance is assessed through a filtering process governed by the classifier's ability to accurately classify it as belonging to the minority class. This step ensures the utility and quality of the synthetic data.

The algorithm proceeds to train the classifier using the balanced and filtered training set, subsequently testing its predictive prowess on the test set. The final step involves reporting and analyzing performance metrics such as the accuracy, the precision and the G-Mean and providing insights into the efficacy of the classifier post-application of the filtering extension. The proposed algorithm's novelty rests on its filtering extension, which discerns the quality of synthetic instances, thereby providing a controlled and refined approach to the SMOTE technique. This ensures that the synthetic data contributes meaningfully to the machine learning model's ability to generalise and accurately predict minority class instances, as shown in Figure 2.

3.1 Dataset

The KEEL (https://sci2s.ugr.es/keel/imbalanced.php) imbalanced dataset repository is a specialized and comprehensive collection designed explicitly for the study and evaluation of machine learning algorithms in the context of imbalanced datasets. It encompasses a wide variety of diverse datasets drawn from different domains, with varying numbers of instances, features, and degrees of class imbalance, thereby providing a robust and challenging environment for testing and validating approaches such as the Machine Learning Based SMOTE Filtering Extension. The diverse nature of the datasets, ranging from slight to severe class imbalances, makes the KEEL repository an invaluable resource for researchers and practitioners working on developing and fine-tuning machine learning models that can effectively handle scenarios where one class is significantly underrepresented compared to others.

3.2 Pre-processing

In this research, the approach diverges from the conventional rule-based filtering of synthetic data. The filtration process hinges on classifying each synthetic instance that belongs to the majority class. To accomplish this, a classifier is trained using the instances adjacent to each minority class instance, which is then utilised to determine the classification of the newly created synthetic instances.
The approach is carried out in eight steps. In the first step, the data set is normalized. The normalisation process, fundamental to this stage, ensures all numerical features are transformed to a common scale, typically between 0 and 1. The mathematical basis for this normalisation often involves a min-max scaling equation (1).

$$X' = \frac{X-Xmi}{Xmax-Xmi} \qquad (1)$$

In the second step, select random sample x from the minority class. In the third step, k neighbors from the minority class around x are selected using Equation [2]. The fourth step selects a random sample x’ from x sample k neighbors. In the fifth step, a new synthetic sample xs is formulated by interpolating between x and its neighbor x’, selecting an intermediate point along the line segment that joins x and x’. Each step, from normalization to distance calculation, is carefully calculated to maintain data integrity and maximize the classifier's efficacy.

$$K(x,x') = \exp\left(-\frac{\parallel x-x' \parallel^2}{2\sigma^2}\right) \qquad (3)$$

where: K: redial base kernel similarity function, x and x′: two data samplesσ: Region of similarity bandwidth gamma.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

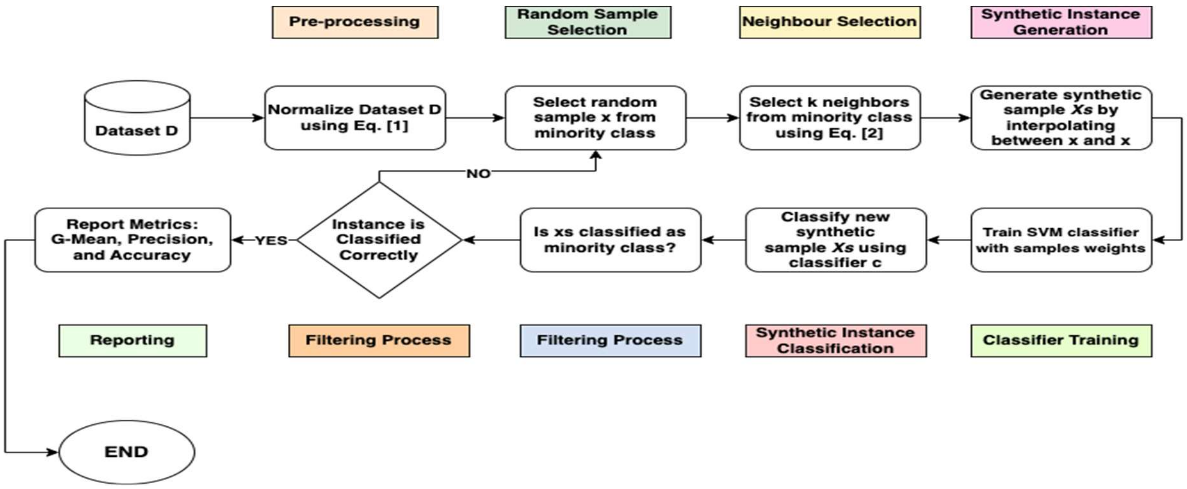
Pre-processing
Random Sample Selection
Neighbour Selection
Synthetic Instance Generation
Dataset D
Normalize Dataset D using Eq. [1]
Select random sample x from minority class
Select k neighbors from minority class using Eq. [2]
Generate synthetic sample Xs by interpolating between x and x
NO
Report Metrics: G-Mean, Precision, and Accuracy
YES
Instance is Classified Correctly
Is xs classified as minority class?
Classify new synthetic sample Xs using classifier c
Train SVM classifier with samples weights
Reporting
Filtering Process
Filtering Process
Synthetic Instance Classification
Classifier Training
END

Figure 3: Flowchart for the Method

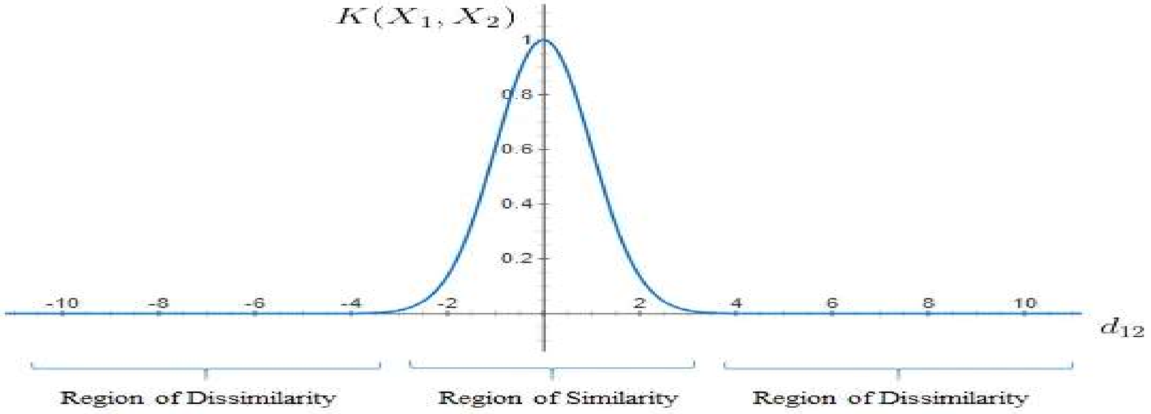

Figure 4: Radial Base Kernel Similarity Function

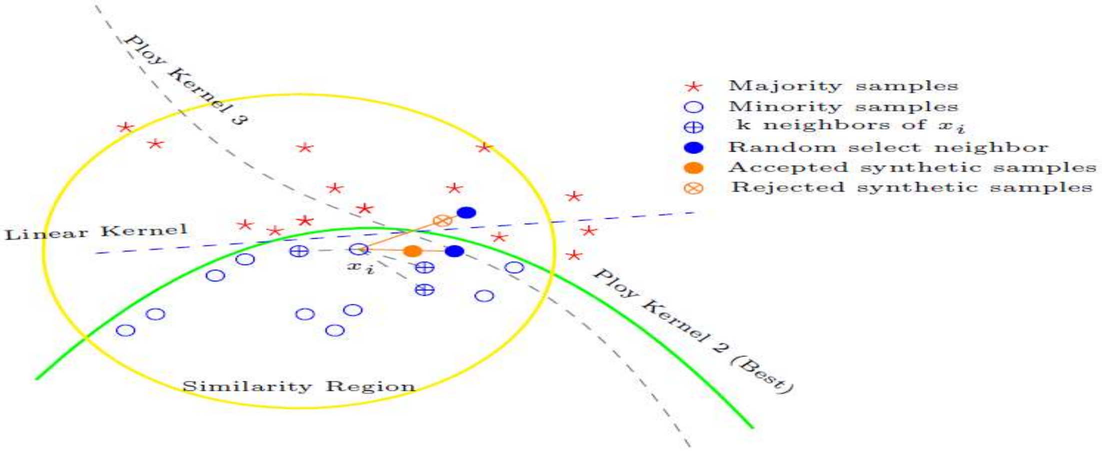

Figure 5: Example of Synthetic Samples

3.3 Mechanism

The research methodology is meticulously structured in eight steps, each designed to enhance the dataset's balance and quality. Following the normalisation of the dataset, the process involves selecting a random minority class sample and identifying its k nearest neighbours using a specific equation. The subsequent steps include interpolating between a selected sample and its neighbour to generate a synthetic sample, and then training an SVM classifier with weighted samples. This classifier then predicts the class of the synthetic sample. Based on this prediction, the synthetic sample is either added to the dataset or discarded, depending on whether it falls within a safe or noisy region. This cycle repeats until an optimal balance between the classes is achieved, ensuring a refined and representative dataset for training the machine learning model. as shown in Figure 2.

The classifier then predicts the class of each synthetic sample. Samples correctly identified as minority class are added to the dataset, with the process repeating until class balance is achieved. This methodology, while intricate, ensures the generation of high-quality synthetic data for improved model training.

Algorithm 1: Machine Learning Based SMOTE Filtering Extension

```
Input: dataset D, learning classifier c, majority class nearest neghbors k
Output: Accuracy, Precision and G-Mean
/* pre-processing
Normalized dataset D using Eq.[1]
Split the dateset $D$ to $D_{tr}$ and testing dataset $D_{ts}$
Calculate the Euclidian distance between each instance in $D_{tr}$ using Eq. [2]
while $D_{tr}$ is not balanceddo
Select random instance in minority class x
    Find k nearest neighbors for x from minority class
    Select random neighbor $x'$from k nearest neighbors
    Interpolate $x^s$ in random distance between x and $x'$
    for each sample $x_i$ in $D_{tr}$do
    Assigning sample weights to base on its similarity with x using Eq. [3]
    end for
    train c on $D_{tr}$ using assigned weights
if $x^s$ classified using c as majoritythen
/* $x^s$ in safe region
            $D_{tr} \leftarrow D_{tr} U x^s$
    Else
        /* $x^s$ in noise region
    /* discard
    end if
end while
/* training
train c using $D_{tr}$
/* testing
for each sample x in $D_{tr}$do
Predict x ucing c
end for
/* reporting
Report the metrics: Accuracy, Precision and G-Mean
```

3.4 Experimental Settings

In the experimental setup for the proposed study, a rigorous testing framework was established. The experiment utilized a 10-fold cross-validation method, ensuring a comprehensive evaluation of the SVM classifier's performance. The dataset was divided, with 80% allocated for training and 20% reserved for testing, offering a substantial data pool for both training the classifier and assessing its predictive capabilities. A critical aspect of this setup was using SVM as the machine learning classifier, specifically focusing on identifying the five nearest neighbours for each minority class instance. This setup was designed to rigorously test the classifier's effectiveness under realistic, varied conditions.

Table 1: Experimental Setting

Parameter	Setting
Number of cross-validation folds	10
Training data percent	80%
Testing data percent	20%
Machine Learning Classifer	SVM
minority class nearest neighbors	5

3.5 Performance Metrics

The performance of the proposed model is measured by the G-mean and AUC metrics. The G-Mean metric is particularly informative for imbalanced datasets as it provides a geometric mean of sensitivity and specificity, offering insight into the balance between class predictions. The AUC, or Area Under the Receiver Operating Characteristic Curve, measures the model's ability to distinguish between classes across different thresholds. High values in both metrics indicate a model that performs well not just on the majority class but is also sensitive to the minority class, encapsulating the essence of a well-balanced classifier. These metrics are reported to demonstrate the classifier's performance comprehensively, highlighting the efficacy of the synthetic data generation and filtering methodology.

4. RESULTS AND DISCUSSION

The results presented in Table 2 and 3 offer a comprehensive view of the performance of the Machine Learning Based SMOTE Filtering Extension algorithm.

4.1 Accuracy Analysis

Table 1 shows that the proposed SMOTE method presents notable enhancements in accuracy in most datasets, with the accuracy improvement from 79.15% to 81.23%. This is exemplified in datasets like 'ecoli4' and 'glass0', where the increases are from 91.27% to 97.13%, and from 57.03% to 67.03%, respectively. However, in few datasets like ecoli1, haberman showed a little decline in the performance. Overall, the proposed SMOTE method has improved the accuracy in 14 out of 20 datasets, indicating its robustness in classifying imbalanced data against other methods [1].

Table 2: Detailed Accuracy Scores for Different Datasets

Name	RC-SMOTE	Our SMOTE
ecoli1	(89.24)	88.19
ecoli2	95.80	(95.88)
ecoli3	86.05	(89.78)
ecoli4	91.27	(97.13)
glass0	57.03	(67.03)
glass2	59.61	(76.11)
glass4	86.52	(87.11)
glass6	86.46	(88.84)
haberman	(74.44)	74.25
new-thyroid1	89.81	(93.19)
Pima	73.41	(73.95)
segment0	91.00	(93.65)
vehicle0	(83.23)	83.23
vehicle1	68.82	(69.24)
vehicle3	68.36	(69.44)
vowel0	87.52	(90.23)
wisconsin	(96.80)	94.81
yeast1-7	(88.43)	86.11
yeast2-4	94.28	(95.12)
yeast3	(94.10)	92.45
Avg.	79.15	81.23
Count	6	14

The comparison between various datasets is illustrated in Figure 6. The proposed filter of SMOTE has demonstrated a superior performances in dealing with the imbalanced data comparing to the RC-SMOTE benchmark method [1], It demonstrated its robustness and reliability for improving classification accuracy. It provides a consistent accuracy improvement, particularly in imbalanced datasets, making it a valuable approach for improving model performance in challenging classification tasks. However, in some occasion it shows a nominal improvement.

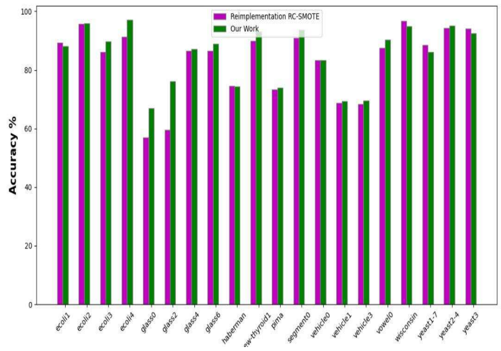

Figure 6: Accuracy Comparison among Datasets

4.2 Precision Analysis

With respect to the precision, Table 3 shows that the proposed SMOTE method outperforming RC-SMOTE in the majority of cases, with an average increase from 70.55% to 72.78%. The proposed method has improved the performance in 13 out of 20 datasets, suggesting a more reliable prediction and a reduction in false positive rates, which is essential in applications where the precision of prediction is critical.

The proposed SMOTE method achieves a 2.23% average increase in precision, improving from 70.55% to 72.78%. This shows that SMOTE enhances the model’s ability to make more accurate positive predictions, which is important when the majority class dominates and skews model predictions. By increasing precision, the SMOTE method helps reduce false positives, which occur when the model incorrectly labels a negative instance as positive. This is especially important the applications such as medical diagnosis, where false positives can result in unnecessary tests or treatments, or in fraud detection, where false positives can waste resources by flagging legitimate transactions as fraudulent.

Table 3: Detailed Precision Scores for Different Datasets

Name	RC-SMOTE	Our SMOTE
ecoli1	(84.47)	83.21
ecoli2	91.00	(91.56)
ecoli3	70.78	(75.20)
ecoli4	72.15	(87.92)
glass0	(71.72)	69.71
glass2	50.19	(65.17)
glass4	(65.42)	61.93
glass6	72.23	(77.71)
haberman	65.98	(66.15)
new-thyroid1	82.60	(89.77)
Pima	71.14	(71.46)
segment0	(80.74)	74.31
vehicle0	(77.32)	77.32
vehicle1	64.78	(71.23)
vehicle3	64.23	(66.12)
vowel0	71.13	(76.13)
wisconsin	(95.92)	93.14
yeast1-7	(62.86)	61.44
yeast2-4	83.51	(84.34)
yeast3	83.45	(84.66)
Avg.	70.55	72.78
Count	7	13

The proposed SMOTE method outperforms RC-SMOTE in 13 out of 20 datasets (Figure 7), showing it is more effective at reducing false positives and improving model reliability. This suggests that SMOTE works well across various datasets, even with different levels of imbalance. This suggests that the SMOTE method generates better synthetic samples that better represent the true distribution of data, leading to more accurate positive predictions and fewer misclassifications.

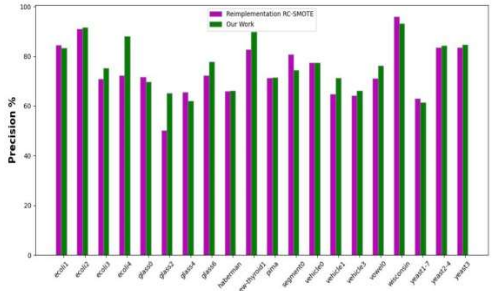

Figure 7: Precision Comparison Among Datasets

4.3 G-Mean Analysis

In Table 3, the G-Mean scores showcase that our proposed SMOTE method has consistent improvement over RC-SMOTE, with an average increase from 76.01% to 78.17%, reflecting a balanced performance between sensitivity and specificity. This is particularly important in the context of imbalanced datasets where ensuring a fair representation of both minority and majority classes is crucial. Moreover, the proposed SMOTE method outperforms in 14 out of 20 datasets, demonstrating its efficacy in handling the nuanced challenges of multi-class imbalanced datasets. The method appears to achieve a commendable trade-off between accurately identifying class labels and maintaining the performance balance between classes, positioning it as a valuable approach in the domain of imbalanced data classification. The G-Mean metric, which is key for evaluating models on imbalanced datasets, is used to assess how well the proposed SMOTE method balances sensitivity and specificity.

Table 4: Detailed G-Mean Scores for Different Datasets

Name	RC-SMOTE	Our SMOTE
ecoli1	(89.02)	88.62
ecoli2	(94.21)	93.91
ecoli3	85.27	(88.41)
ecoli4	(92.37)	88.28
glass0	59.74	(77.12)
glass2	54.25	(69.16)
glass4	85.05	(85.67)
glass6	82.93	(87.25)
Haberman	56.52	(58.50)
new-thyroid1	90.17	(91.44)
Pima	(71.58)	70.47
segment0	90.99	(93.51)
vehicle0	(79.07)	79.07

vehicle1	68.39	(70.11)
vehicle3	68.06	(69.16)
vowel0	91.60	(92.13)
wisconsin	(97.39)	95.35
yeast1-7	61.26	(64.13)
yeast2-4	87.44	(88.26)
yeast3	90.94	(91.12)
Avg.	76.01	78.17
Count	6	14

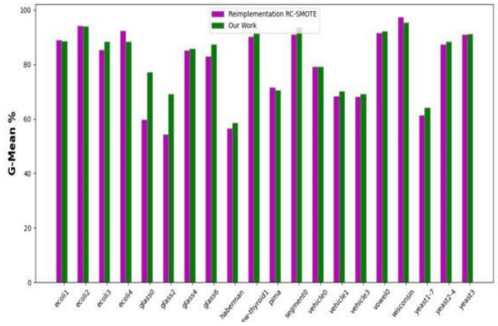

Figure 9: G-Mean Score Comparison among Datasets

5. CONCLUSION

This research proposed a novel methodological framework that integrates machine learning-based SMOTE filtering extension algorithm, the k-nearest neighbor and the tailored kernel applications with support vector machines (SVM). This technique has demonstrated significant advancements in addressing imbalanced datasets and overlapping while reducing bias towards majority classes. The enhanced algorithm, with its structured approach to synthetic data generation and refined classification methodology, has shown promising results in G-Mean, AUC, accuracy, and precision metrics across various datasets. The proposed SMOTE technique shows a notable enhancement comparing to other methods in terms of accuracy from 79.15% to 81.23%, precision from 70.55% to 72.78% and G-mean from 76.01% to 78.17% based-on several datasets. These experimental outcomes underscore the method's effectiveness and robustness in not only improving class balance but also in boosting classifier performance in more complex multi-class scenarios. However, this study also paved the way to the future research to explore more robust, innovative and powerful technique for the practical applicability considering more diverse datasets to cater to even rarer and more complex diseases. This extended analysis solidifies the potential of this innovative approach, marking a substantial progression in the field of imbalanced data classification.

REFERENCES

[1] Soltanzadeh, P., & Hashemzadeh, M. (2021). RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. Information Sciences, 542, 92–111. https://doi.org/10.1016/j.ins.2020.07.014.

[2] Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on Imbalance Medical Data. IEEE Access, 9, 109960–109975. https://doi.org/10.1109/access.2021.3102399.

[3] Lin, C., Tsai, C.-F., & Lin, W.-C. (2022). Towards hybrid over- and under-sampling combination methods for class imbalanced datasets: An experimental study. Artificial Intelligence Review, 56(2), 845–863. https://doi.org/10.1007/s10462-022-10186-5.

[4] Li, H. (2023). Support Vector Machine. Machine Learning Methods, 127–177. https://doi.org/10.1007/978-981-99-3917-6_7.

[5] Mittal, N., Pandit, A. K., Abouhawwash, M., & Mahajan, S. (2024). Intelligent Systems and applications in Computer Vision. CRC Press, Taylor and Francis Group.

[6] Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). Smote–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Information Sciences, 291, 184–203. https://doi.org/10.1016/j.ins.2014.08.05.

[7] García, S., Zhang, Z.-L., Altalhi, A., Alshomrani, S., & Herrera, F. (2018). Dynamic Ensemble selection for multi-class imbalanced datasets. Information Sciences, 445–446, 22–37. https://doi.org/10.1016/j.ins.2018.03.002.

[8] Mehmood, Z., & Asghar, S. (2021). Customizing SVM as a base learner with AdaBoost ensemble to learn from multi-class problems: A hybrid approach AdaBoost-MSVM. Knowledge-Based Systems, 217, 106845. https://doi.org/10.1016/j.knosys.2021.106845.

[9] Rossi, S., Acampora, G., & Staffa, M. (2020). Working together : a DBN approach for individual and group activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, *11*(12), 6007–6019. https://doi.org/10.1007/s12652-020-01851-0

[10] Luo, F., Khan, S., Huang, Y., & Wu, K. (2021). Binarized neural network for edge intelligence of sensor-based human activity recognition. *IEEE Transactions on Mobile Computing*, *1233*(c), 1–13. https://doi.org/10.1109/TMC.2021.3109940

[11] Lentzas, A., Dalagdi, E., & Vrakas, D. (2022). Multilabel Classification Methods for Human Activity Recognition: A Comparison of Algorithms. *Sensors*, *22*(6). https://doi.org/10.3390/s22062353

[12] Jethanandani, M., Sharma, A., Perumal, T., & Chang, J. R. (2020). Multi-label classification based ensemble learning for human activity recognition in smart home. *Internet of Things (Netherlands)*, *12*, 100324. https://doi.org/10.1016/j.iot.2020.100324

[13] Hamad, R. A., Kimura, M., & Lundström, J. (2020). Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments. *SN Computer Science*, *1*(4), 1–10. https://doi.org/10.1007/s42979-020-00211-1

[14] Charte, F., Rivera, A. J., Del Jesus, M. J., & Herrera, F. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, *89*, 385–397. https://doi.org/10.1016/j.knosys.2015.07.019.

[15] Abokadr, S., Azman, A., Hamdan, H., & Amelina, N. (2023). Enhancing rare disease diagnosis: a weighted cosine similarity approach for improved k-nearest neighbor algorithm. Journal of Theoretical and Applied Information Technology, 101(17).