An Approach for Improving Oversampling by Filtering out Unrealistic Synthetic Data

Nada Boudegzdame[1][a], Karima Sedki[1][b], Rosy Tspora[2,3,4][c] and Jean-Baptiste Lamy[1][d]

[1]*LIMICS, INSERM, Université Sorbonne Paris Nord, Sorbonne Université, France*

[2]*INSERM, Université de Paris Cité, Sorbonne Université, Cordeliers Research Center, France*

[3]*HeKA, INRIA, France*

[4]*Department of Medical Informatics, Hôpital Européen Georges-Pompidou, AP-HP, France*

Keywords: Imbalanced Data, Oversampling, SMOTE, Class Imbalance, Data Augmentation, Machine Learning, Neural Networks, Synthetic Data, Synthetic Sample Detector, Generative Adversarial Networks.

Abstract: Oversampling algorithms are commonly used in machine learning to address class imbalance by generating new synthetic samples of the minority class. While oversampling can improve classification models' performance on minority classes, our research reveals that models often learn to detect noise generated by oversampling algorithms rather than the underlying patterns. To overcome this issue, this article proposes a method that involves identifying and filtering unrealistic synthetic data, using advanced technique such a neural network for detecting unrealistic synthetic data samples. This aims to enhance the quality of the oversampled datasets and improve machine learning models' ability to uncover genuine patterns. The effectiveness of the proposed approach is thoroughly examined and evaluated, demonstrating enhanced model performance.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Class imbalance is a common challenge in machine learning, occurring when one class has significantly fewer samples than others. To tackle this issue, oversampling techniques, such as the Synthetic Minority Over-sampling Technique (*SMOTE*) (Chawla et al., 2002), have been widely employed.

Oversampling, while enhancing model performance on minority classes, presents challenges, especially in highly imbalanced datasets. In such cases, the resulting oversampled dataset for the minority class is predominantly synthetic, overshadowing the original data. This dominance may lead the model to prioritize predicting the synthetic nature by capturing noise introduced during oversampling, rather than discerning the genuine underlying patterns. Consequently, it can result in poor generalization and suboptimal real-world performance (Tarawneh et al., 2022; Drummond and Holte, 2003; Chen et al., 2004; Rodríguez-Torres et al., 2022).

This article proposes a methodological approach to improve synthetic data quality by training a machine learning model to predict the synthetic status of each sample. The goal is to identify and filter unrealistic synthetic data, thereby improving overall dataset quality and enhancing the model's ability to uncover genuine underlying patterns. Our study comprehensively investigates the proposed approach's performance on diverse datasets, focusing on its effectiveness in improving synthetic data quality and enhancing machine learning model performance on oversampled data. The research aims to contribute effective strategies for handling class imbalance and overcoming detectability issues associated with synthetic data.

2 BACKGROUND

Various oversampling techniques address class imbalance, with *SMOTE* being a prominent method (Chawla et al., 2002). It interpolates between minority samples to generate new ones along the connecting line. Over the years, *SMOTE* has undergone numerous modifications and extensions to enhance its effectiveness, addressing issues such as overfitting, data

[a] https://orcid.org/0000-0003-1409-6560
[b] https://orcid.org/0000-0002-2712-5431
[c] https://orcid.org/0000-0002-9406-5547
[d] https://orcid.org/0000-0002-5477-180X

Boudegzdame, N., Sedki, K., Tspora, R. and Lamy, J.
An Approach for Improving Oversampling by Filtering out Unrealistic Synthetic Data.
DOI: 10.5220/0012325400003636
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 291-298
ISBN: 978-989-758-680-4; ISSN: 2184-433X
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

density, and mixed feature types.

Enhancements to *SMOTE* include *Borderline SMOTE* (Han et al., 2005), focusing on addressing overfitting by generating synthetic samples near the decision boundary. *ADASYN* (He et al., 2008) adjusts the density distribution to generate more samples for harder-to-learn instances, while *Safe-Level SMOTE* (Bunkhumpornpat et al., 2009) reduces misclassification risks by focusing on samples near a safe majority class. *SMOTEN* (Chawla et al., 2002) extends *SMOTE* for datasets with mixed nominal and continuous features, employing a tailored distance metric for generating synthetic samples specifically for nominal features. Additionally, *Minority Oversampling Technique (MOTE)* (Huang et al., 2006) generates synthetic samples exclusively for those misclassified by the current model.

Choosing an oversampling method requires thoughtful consideration due to differing strengths and weaknesses. The chosen technique significantly influences the model's performance, emphasizing the need for a methodological approach to address detectability issues and enhance synthetic data quality.

In addition to oversampling techniques, the emergence of generative adversarial networks (GANs) (Goodfellow et al., 2014) offer alternative methods for synthetic data generation. GANs employ a competitive training approach, where two neural networks are trained jointly: one network generates realistic synthetic data, while the other network discriminates between real and synthetic data. They have demonstrated success in generating complex synthetic data, such as images and text (Mirza and Osindero, 2014; Reed et al., 2016; Zhang et al., 2017).

Despite their success in generating realistic data, GANs struggle with categorical synthetic datasets due to gradient computation limitations on latent categorical variables. Methods like medGAN (Choi et al., 2017), which transforms categorical data using autoencoders, have been developed to address this limitation. However, medGAN is limited to binary and count data, leading to the development of MC-MedGAN (Camino et al., 2018) for multi-categorical variables.

Our generic method, applicable alongside any oversampling technique, aims to enhance synthetic data quality without relying on an internal generative component. Incorporating our approach into the oversampling process provides a flexible and effective solution for improving model performance on imbalanced datasets.

3 CHALLENGES OF OVERSAMPLING

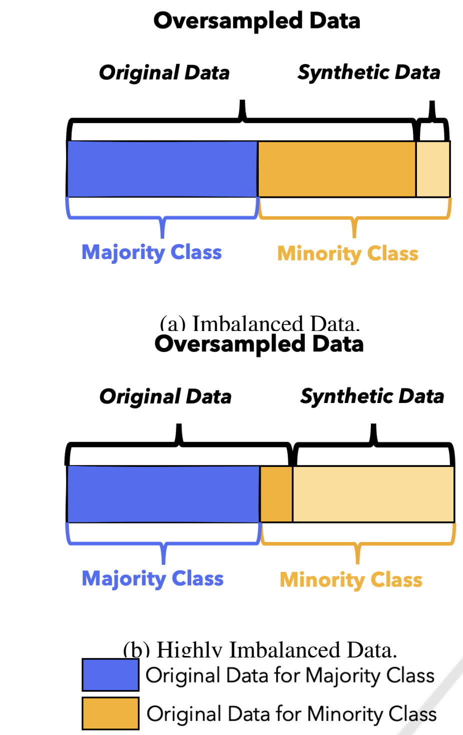
Intuitively, an effective oversampling technique increases the representation of the minority class without merely replicating existing instances. For instance, *SMOTE* generates synthetic instances by interpolating between existing minority class instances and their k nearest neighbors. However, oversampling, while enhancing model performance on minority classes, introduces significant challenges.

Firstly, oversampling can induce bias towards the minority class, causing the model to prioritize it at the expense of the majority class. This bias can lead to poor performance on real-world data, where the minority class is less frequent (Tarawneh et al., 2022; Drummond and Holte, 2003). Another issue may arise from potential inconsistencies in data types, as synthetic data points may deviate from the typical range or adopt different formats.

Moreover, oversampling is prone to generating mislabeled samples belonging to the majority class or creating unrealistic ”noise” samples. It can alter the data distribution, impacting the representation of different class proportions. Additionally, oversampling may reduce dataset diversity, potentially leading to overfitting and hindering generalization to new data by creating synthetic samples closely resembling existing ones.

A careful assessment of oversampling's impact on dataset distribution and diversity is essential to ensure the resulting model accurately reflects the true nature of the problem. Additionally, consideration of the computational costs associated with generating synthetic data is crucial, especially for large datasets, given the time-consuming and resource-intensive nature of the process (Chen et al., 2004; Rodríguez-Torres et al., 2022).

To gain a more profound understanding of why machine learning models often lean towards learning the synthetic nature of data over genuine underlying patterns, we examine the data distribution in oversampled data. In slightly imbalanced data, depicted in Figure 1a, where only a few synthetic samples are required for class balance, the overall class distribution remains largely unchanged. In highly imbalanced datasets, oversampling becomes particularly challenging, given that the minority class constitutes a very small fraction of the data. The substantial number of synthetic instances required to balance the data results in the majority of the oversampled data being synthetic, while the original data makes up only a small fraction. As illustrated in Figure 1b, the majority class tends to be equivalent to the original data



(a) Imbalanced Data.

(b) Highly Imbalanced Data.

Figure 1: Class Distribution in Oversampled Dataset.

class, while the minority class classification tends to be equivalent to the synthetic data class, as the original minority class samples contribute negligibly to the data for that class.

Hence, the quality of synthetic data significantly impacts machine learning model performance, especially for the minority class. It is crucial for synthetic data to accurately mirror real-world data. Inclusion of unrealistic synthetic data may lead the model to misclassify the minority class as synthetic, causing it to predict these instances rather than capturing genuine underlying patterns. This situation results in a redefinition of the learning problem, shifting the focus to predicting the synthetic nature of the data.

Addressing the challenge of unrealistic synthetic data is paramount for enhancing the model's ability to discern and leverage essential patterns, ultimately improving performance on real-world datasets. This is particularly crucial in highly imbalanced scenarios where there is an increased likelihood of the model may detecting the synthetic nature. Therefore, generating realistic synthetic data is vital to mitigate this issue and ensure overall dataset quality.

4 METHOD

Our solution to improve the quality of synthetic data is iterative and consists of three main steps, as illustrated in Figure 2. First, we generate synthetic data using the chosen oversampling technique. The second step, which occurs only during the initial iteration, involves building a machine learning model trained to predict the synthetic status of each sample in the dataset. In the third step, this model is then employed to identify and filter out **detectable** synthetic data. The predictive model is used to flag and remove samples classified as synthetic and unrealistic.

By eliminating these **unrealistic** synthetic samples, our aim is to enhance the overall quality of the dataset and mitigate the negative impact of the noise introduced by these samples, enabling the machine learning model to focus on the genuine underlying patterns in the original data. The following subsections provide a detailed explanation of each step.

4.1 Generation of Synthetic Data

The first step of the proposed method consists of generating synthetic data from the minority class to balance the overall distribution of classes in the data. For this step, we can use any existing oversampling technique. As our method can be adapted to various oversampling techniques, it is very flexible which is useful as the choice of the most appropriate technique depends on the dataset.

4.2 Learning the Detectability of Synthetic Data

In the second step, we aim to assess the detectability of synthetic data and identify unrealistic instances for subsequent filtering. To achieve this, we formulate a binary classification problem to distinguish between synthetic and original data samples based on their distinct characteristics. We first prepare the dataset for the learning phase:

1. **Build the Dataset:** Remove samples from the majority class in the oversampled dataset generated in *STEP 1*, as the focus is on detecting the synthetic nature of data generated from the minority class.
2. **Label the Samples:** Assign labels to each instance in the refined dataset, indicating whether it is synthetic (1) or original (0).

This binary classification problem, summarized bellow, aims to train a machine learning model to

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

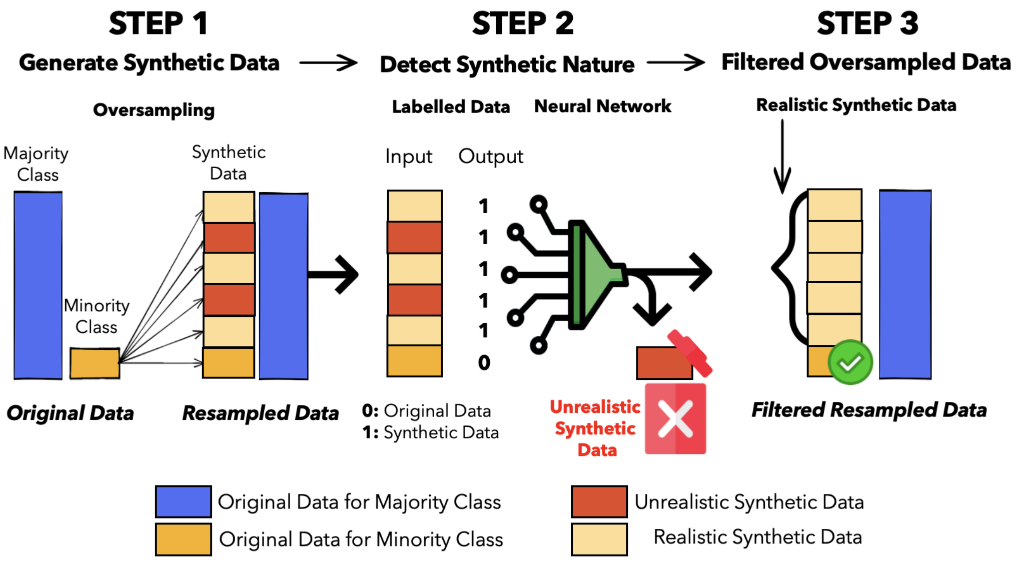


Figure 2: Illustration of the proposed oversampling filtering technique.

distinguish between synthetic and original instances based on their distinctive characteristics. By capturing the underlying patterns and features that differentiate synthetic data from original data, the model could learn to predict the synthetic status of each sample.

Synthetic Sample Detector
Input: *Oversampled data: original and synthetic data*
Output: *Is the instance synthetic or original?*

This problem formulation served as the foundation for exploring the detectability of synthetic data and identifying lower-quality instances. The insights gained from this analysis played a crucial role in guiding the subsequent step of filtering out detectable synthetic data, which aimed to enhance the overall quality of the synthetic dataset and improve the performance of machine learning models on imbalanced datasets.

The second step is only performed at the first iteration. In further iterations, we reuse the same model generated at the first iteration without learning a new model.

4.3 Filtering out Unrealistic Synthetic Data

In the final step, we employ the synthetic data detector created in *Step 2* to predict the synthetic nature of each data instance generated in *Step 1*. Instances identified as synthetic data are filtered out, retaining only those closely resembling the characteristics of the original data. If the remaining data remains imbalanced, we iteratively generate additional synthetic data samples, detecting and filtering out unrealistic synthetic data in each iteration. This process continues until achieving a desired balance between the minority and majority classes. It facilitates the progressive enhancement of synthetic data quality, diminishing the detectability of synthetic instances, and consequently, improving the model's accuracy in predicting the minority class.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

5 EXPERIMENTS

5.1 Experimental Databases

In this experimental study, we assessed our data filtering technique on oversampled data from various techniques, including *SMOTE*, *Borderline SMOTE*, *SMOTEN* , *ADASYN*, across diverse databases. These databases, carefully selected for their diversity, represent different real-world problems with varying class imbalances and domains:

- **Credit Card Fraud:** Highly imbalanced (0.17% minority class ratio); excellent representation of real-world financial transactions with infrequent fraudulent activities.
- **Car Insurance Claim:** Moderately imbalanced (6.39% minority class ratio); reflecting imbalances in insurance claims between common and rare cases.
- **Anomalies in Wafer Manufacturing:** Intermediate imbalance (8.11% minority class ratio); mimics manufacturing scenarios where detecting anomalies is essential.

Table 1: Database descriptions.

Database	Domain	Input types	Feature number	Minority Class Ratio
Haemorrhage (MIMIC)	Medical	Boolean	5317	3.00 %
Credit Card Fraud	Finance	Real	29	0.17 %
Student Dropout	Education	Boolean, Real	36	32.12 %
Anomalies in Wafer M.	Manufacturing	Boolean, Real	1558	8.11 %
Car Insurance Claim	Insurance	Boolean, Real	42	6.39 %

- **Haemorrhage (MIMIC):** Significantly imbalanced (3% minority class ratio); representing haemorrhage risk and non-risk instances using MIMIC database, mirroring real-world medical scenarios where certain conditions are infrequent.
- **Student Dropout:** Relatively balanced yet still imbalanced (32.12% minority class ratio); pertaining to the education sector where dropout events are infrequent compared to student persistence.

These databases cover diverse domains: finance, insurance, manufacturing, medical, and education. They reflect real-world situations where rare events or anomalies occur less frequently than normal instances. By evaluating the filtering technique across databases with varying minority class ratios, we gained insights into its ability to effectively identify and filter out unrealistic synthetic data. This evaluation allowed us to assess the technique's generalizability and applicability across different real-world scenarios.

Table 1 provides a summary of the databases used in our experiments, including their respective domains, input types, feature numbers, and minority class ratios.

5.2 Implementation and Performance Metrics

To assess the performance of our approach, we employed a neural network with a tailored architecture for each dataset. The network used *LeakyReLU* activation functions to address ”dead” neurons and employed a *sigmoid* activation function for the output layer, well-suited for binary classification tasks. We optimized training using the ***ReduceLROnPlateau*** technique, dynamically adjusting the learning rate to enhance efficiency and prevent convergence plateaus.

During evaluation, we considered precision, recall, and F1 score to assess the model's ability to correctly identify the minority class (He and Garcia, 2009; Powers, 2011). These metrics collectively offer a comprehensive understanding of the model's strengths and limitations, ensuring a holistic assessment. Relying on a single metric cannot provide a complete evaluation. These metrics provided valuable insights into the model's ability to address imbalanced class distributions and improve the overall quality of oversampled data.

Our design choices and selected metrics aimed to enhance the neural network's learning capability and overall effectiveness. The next section presents experimental results, highlighting our technique's performance across diverse datasets and varying minority class ratios.The metrics, together with information about the imbalance levels in the datasets, will offer insights into how effectively the technique addresses imbalanced class distributions.

6 RESULTS

In our experimental study, we rigorously evaluated the proposed Filtering Oversampling method across diverse learning problems. We followed a systematic approach, assessing machine learning models on original datasets as a baseline, then evaluating oversampled data using various techniques. The final step involved testing models on filtered oversampled data, enabling a direct comparison with popular oversampling techniques (*SMOTE*, *Borderline SMOTE*, *SMOTEN*, *ADASYN*). Summary performance metrics for each experiment are presented in Table 2.

The Filtering Oversampling method with *SMOTE* demonstrated significant improvements in predicting haemorrhage risk. The f1 score increased from 0.62 to 0.69, recall improved from 0.60 to 0.82, and accuracy rose from 0.63 to 0.88. In contrast, the standalone use of *SMOTE* without filtering resulted in a much lower F1 score of 0.12, recall of 0.21, precision of 0.09, and accuracy of 0.88, indicating poor real-world generalization. Other oversampling techniques, such as Borderline *SMOTE*, *SMOTEN*, and *ADASYN*, showed varying degrees of effectiveness but fell short of the enhancements achieved by the Filtering Oversampling method, as a substantial portion of all the possible synthetic data that can be generated were detectable, resulting in an incomplete balance of the

Table 2: Performance Metrics Comparison of the Filtering Oversampling Method.

Learning Problem	**Oversampled**	**Filtered**	**F1 score**	**Recall**	**Precision**	**Accuracy**
Haemorrhage Risk Prediction	No	-	0.62	0.60	**0.65**	0.63
	SMOTE	no	0.12	0.21	0.09	0.88
	SMOTE	yes	**0.69**	**0.82**	0.60	0.88
	Borderline SMOTE	no	0.05	0.04	0.09	**0.94**
	Borderline SMOTE	yes	0.15	0.23	0.11	0.91
	SMOTEN	no	0.05	0.21	0.09	0.92
	SMOTEN	yes	0.11	0.35	0.06	0.80
	ADASYN	no	0.09	0.11	0.07	0.90
	ADASYN	yes	0.17	0.19	0.15	0.92
Credit Card Fraud Detection	No	-	0.23	0.77	0.13	**0.99**
	SMOTE	no	0.0042	0.22	0.0021	0.83
	SMOTE	yes	**0.91**	0.84	**1.0**	0.84
	Borderline SMOTE	no	0.23	0.90	0.13	**0.99**
	Borderline SMOTE	yes	0.62	0.81	0.5	0.98
	SMOTEN	no	0.04	0.86	0.02	0.94
	SMOTEN	yes	0.19	0.86	0.10	0.98
	ADASYN	no	0.08	**0.94**	0.04	0.96
	ADASYN	yes	0.57	0.82	0.44	**0.99**
Student Dropout Prediction	No	-	0.61	0.51	0.74	0.73
	SMOTE	no	0.72	0.73	0.71	0.77
	SMOTE	yes	0.78	0.70	0.88	0.84
	Borderline SMOTE	no	0.80	0.74	0.86	0.85
	Borderline SMOTE	yes	**0.85**	**0.86**	0.85	**0.88**
	SMOTEN	no	0.77	0.74	0.81	0.84
	SMOTEN	yes	0.80	0.81	0.74	0.83
	ADASYN	no	0.79	0.70	**0.91**	0.86
	ADASYN	yes	0.83	0.85	0.80	0.86
Detecting Anomalies in Wafer Manufacturing	no	-	0.47	0.75	0.34	0.86
	SMOTE	no	0.50	0.74	0.37	0.89
	SMOTE	yes	0.57	0.51	0.64	0.93
	Borderline SMOTE	no	0.60	0.72	0.51	0.92
	Borderline SMOTE	yes	0.55	0.60	0.51	0.91
	SMOTEN	no	0.52	0.64	0.44	0.88
	SMOTEN	yes	**0.73**	**0.90**	0.61	**0.94**
	ADASYN	no	0.48	0.75	0.35	0.85
	ADASYN	yes	0.68	0.64	**0.72**	0.93
Car Insurance Claim	No	-	0.08	0.11	0.06	**0.83**
	SMOTE	no	0.10	0.44	0.06	0.54
	SMOTE	yes	**0.12**	**0.67**	**0.07**	0.38
	Borderline SMOTE	no	**0.12**	0.66	0.06	0.41
	Borderline SMOTE	yes	0.11	0.60	0.06	0.36
	SMOTEN	no	**0.12**	0.65	**0.07**	0.38
	SMOTEN	yes	0.11	0.47	0.07	0.53
	ADASYN	no	**0.12**	0.61	0.06	0.43
	ADASYN	yes	**0.12**	0.61	**0.07**	0.43

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

dataset.

For credit card fraud detection, the original dataset had limited fraud detection capability (F1 score of 0.23). While *SMOTE* without filtering led to a marginal deterioration (F1 score 0.0042), indicating a redefinition of the learning problem due to the introduced noise. In stark contrast, when combined with Filtering Oversampling, a substantial performance boost was observed, with an F1 score of 0.91, a recall rate of 0.84, and a precision of 1.0, showcasing its potency, especially in critical applications like fraud detection. Other methods (*Borderline SMOTE*, *SMOTEN*, *ADASYN*) exhibited varying degrees of effectiveness.

In the context of predicting student dropout, the utilization of *Borderline SMOTE* initially demonstrated an impressive improvement F1 score from 0.61 to 0.80, reaching its peak effectiveness at 0.85 when combined with our filtering technique. This underscores the effectiveness of integrating oversampling with our filtering method, effectively addressing class imbalance and enhancing predictive accuracy.

In wafer manufacturing anomaly detection, the baseline F1 score is a modest 0.47. Integrating *ADASYN* with our filtering approach leads to a significant improvement, boosting the F1 score from 0.48 to 0.68. *SMOTEN* introduces an initial F1 score of 0.52, and our filtering approach further contributes to a marginal increase, reaching 0.73. These results highlight the effectiveness of our method in enhancing anomaly detection for wafer manufacturing.

However, in car insurance claims, the initial F1 score is 0.08, indicating potential for improvement. Applying *SMOTE* without filtering leads to a slight increase in the F1 score to 0.10, while the integration of our filtering approach with *SMOTE* further boosts the F1 score to 0.12. Additionally, both *SMOTEN* and *Borderline SMOTE* exhibit F1 score of 0.12 and 0.11 with and without filtering, respectively. These findings highlight the variable effectiveness of filtering in enhancing the F1 score, dependent on the oversampling method and dataset.

Overall, our experimental results provide strong evidence that the proposed filtering oversampling method consistently outperforms both the original dataset and the widely used methods (*SMOTE*, *Borderline SMOTE*, *SMOTEN*, *ADASYN*) across a range of learning problems. These findings highlight the effectiveness of our approach in enhancing the performance of imbalanced classification tasks. It's important to acknowledge that high accuracy observed when learning on the original data is primarily influenced by the large number of true negatives. Therefore, it's crucial to consider multiple performance metrics, such as F1 score, precision, and recall, to assess the model's effectiveness in handling imbalanced datasets.

7 DISCUSSION

In this study, we proposed a novel approach to enhance oversampling methods through a filtering mechanism to eliminate unrealistic synthetic data, resulting in substantial performance improvements. Rigorous testing highlights the method's impact on capturing genuine patterns in the minority class, thereby improving generalization and real-world performance. As the model relies less on predicting synthetic instances, it gains robustness to handle challenges in real-world data.

While our method has been applied to well-known oversampling techniques such as *SMOTE*, *Borderline SMOTE*, *SMOTEN*, and *ADASYN*, it is, in essence, a generic approach adaptable to other oversampling techniques. The results showcased that this method excels with highly imbalanced data, which are most impacted by the noise oversampling, given that the majority of the oversampled dataset comprises synthetic instances. The extent of improvement achieved through our filtering method is not quantified by a fixed value; it depends on the dataset, the number, and type of features involved. On small datasets with limited instances and features, the challenge lies in running out of synthetic samples to effectively balance the classes.

It's crucial to note that the model created in *Step 2*, trained on data where the synthetic class may dominate, faces challenges in detecting synthetic data due to class imbalance. Consequently, applying our filtering technique may not significantly improve results in such scenarios. For certain datasets, like the "Car Insurance Claim database" in our experiment, limited improvement was observed due to exhausted synthetic samples and the inability to generate enough realistic synthetic samples (i,e undetected synthetic data) to balance the dataset. This limitation may be attributed to a possible bias toward the unrealistic class, given its majority representation.

In summary, our proposed method provides a valuable contribution to mitigate oversampling technique limitations and enhancing machine learning model performance on imbalanced datasets. By improving synthetic data quality, it enables more accurate learning and better handling of class imbalance in real-world applications.

8 CONCLUSION AND PERSPECTIVES

In conclusion, oversampling techniques provide a valuable approach to address class imbalance in machine learning. Nevertheless, their effectiveness can be hindered by the quality of synthetic data generated during the oversampling process. To overcome this limitation, the proposed filtering oversampling method selectively filters out unrealistic synthetic data, thereby enhancing the performance of machine learning models on imbalanced datasets. This leads to improved performance on real-world datasets as the model becomes less reliant on predicting synthetic instances and gains better generalization capabilities beyond the synthetic data distribution.

For future research, promising directions include incorporating explainability and interpretability aspects into the filtering oversampling method. Developing techniques to understand the impact of filtered synthetic data on the model's decision-making process can enhance insights and prediction trustworthiness. Additionally, extending the research to multi-class classification problems, beyond initial binary classification tasks, will assess the method's effectiveness across a broader range of scenarios.

We aim to advance the understanding and capabilities of handling imbalanced datasets by pursuing these future research directions, ultimately enhancing the performance of machine learning models in real-world applications.

ACKNOWLEDGEMENTS

This work was partially funded by the French National Re- search Agency (ANR) through the ABiMed Project [grant number ANR-20-CE19-0017-02].

REFERENCES

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 475–482.

Camino, R., Hammerschmidt, C., and State, R. (2018). Generating multi-categorical samples with generative adversarial networks. In *ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, pages 1–7.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, 110:24–31.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference*, pages 286–305.

Drummond, C. and Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Neural Information Processing Systems*, pages 2672–2680.

Han, H., Wang, W. Y., and Mao, B. H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

He, H. and Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets.

Powers, D. (2011). Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, volume 48, pages 1060–1069.

Rodríguez-Torres, F., Martínez-Trinidad, J. F., and Carrasco-Ochoa, J. A. (2022). An oversampling method for class imbalance problems on large datasets. *Applied Sciences*, 12(7):3424.

Tarawneh, S., Al-Betar, M. A., and Mirjalili, S. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):340–354.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., and Carin, L. (2017). Adversarial feature matching for text generation. In *International Conference on Machine Learning*, pages 4006–4015.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS