

# SMOTE: Are We Learning to Classify or to Detect Synthetic Data?

Nada Boudegzdame<sup>1</sup><sup>a</sup>, Karima Sedki<sup>1</sup><sup>b</sup>, Rosy Tspora<sup>2,3,4</sup><sup>c</sup> and Jean-Baptiste Lamy<sup>1</sup><sup>d</sup>

<sup>1</sup>LIMICS, INSERM, Université Sorbonne Paris Nord, Sorbonne Université, France

<sup>2</sup>INSERM, Université de Paris Cité, Sorbonne Université, Cordeliers Research Center, France

<sup>3</sup>HeKA, INRIA, France

<sup>4</sup>Department of Medical Informatics, Hôpital Européen Georges-Pompidou, AP-HP, France

**Keywords:** Imbalanced Data, Oversampling, SMOTE, Data Augmentation, Class Imbalance, Machine Learning, Neural Networks, Synthetic Data.

**Abstract:** Oversampling algorithms are used as preprocess in machine learning, in the case of highly imbalanced data in an attempt to balance the number of samples per class, and therefore improve the quality of models learned. While oversampling can be effective in improving the performance of classification models on minority classes, it can also introduce several problems. From our work, it came to light that the models learn to detect the noise added by the oversampling algorithms instead of the underlying patterns. In this article, we will define oversampling, and present the most common techniques, before proposing a method for evaluating oversampling algorithms.

## 1 INTRODUCTION


Oversampling is a technique used to solve the problem of class imbalance in machine learning. Class imbalance occurs when the number of samples in one class is much lower than the number of samples in the other class(es). This is a problem because the classifier will have a hard time learning from the minority class. Oversampling techniques generate additional samples belonging to the minority class so that the classifier has a better chance of learning from them (He and Garcia, 2009; Batista et al., 2004).


Oversampling creates new instances of the minority classes by either 1) replicating existing instances or, 2) synthesizing samples, to increase its representation in the dataset. Some popular techniques include Random Oversampling (Chawla et al., 2002), SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008), Borderline SMOTE (Han et al., 2005), SMOTEN (Chawla et al., 2002), Safe-Level SMOTE (Bunkhumpornpat et al., 2009), and Minority Oversampling Technique (MOTE) (Huang et al., 2006).


While oversampling can enhance the performance


of classification models on minority classes but brings significant problems, especially in highly imbalanced data. In this article, we will define the potential problems and challenges when using oversampling. A core concern arises from the shift in dataset composition due to oversampling, where the original minority class data becomes a small portion, overshadowed by synthetic data. This transformation fundamentally alters the learning problem for machine learning models. Consequently, models often prioritize predicting synthetic data, learning noise instead of underlying minority class patterns. This can result in poor model generalisation and performance on real-world data (Tarawneh et al., 2022; Drummond and Holte, 2003; Chen et al., 2004; Rodríguez-Torres et al., 2022).

Consequently, we propose a method for evaluating oversampling techniques on a given dataset, with a focus on the specific application of drug-induced hemorrhage. The method consists in trying to learn a model that can predict the synthetic status of the sample; the better this model is, the worst the oversampling technique performs. Finally, we will put to test the most common oversampling techniques and evaluate their effectiveness in a practical example.

<sup>a</sup> <https://orcid.org/0000-0003-1409-6560>

<sup>b</sup> <https://orcid.org/0000-0002-2712-5431>

<sup>c</sup> <https://orcid.org/0000-0002-9406-5547>

<sup>d</sup> <https://orcid.org/0000-0002-5477-180X>

## 2 BACKGROUND

Since the introduction of SMOTE over 20 years ago in 2002, numerous techniques have evolved to enhance its effectiveness. Borderline SMOTE was among the first improvements, mitigating SMOTE's overfitting risk by generating synthetic samples exclusively for minority class instances near the decision boundary. ADASYN followed, addressing harder-to-learn minority samples by adapting the density distribution of the feature space. Safe-Level SMOTE selectively generates synthetic samples for minority class instances with proximate majority class neighbors to reduce misclassification risk. The latest advancement, SMOTEN, handles datasets with both nominal and continuous features through a distinct approach for synthetic sample generation in nominal features.

These oversampling techniques exhibit varying strengths and weaknesses, with performance influenced by the dataset and classification task. To further tackle remaining challenges, several approaches have emerged, including combining oversampling with under-sampling, utilizing advanced synthetic sampling techniques, or adjusting classification thresholds. Each approach carries its unique advantages and limitations, necessitating careful selection based on the specific dataset and classification problem at hand.

### 2.1 Known Problems of Oversampling

SMOTE, the most common (He and Garcia, 2009) and effective (Batista et al., 2004) oversampling technique, generates synthetic minority class instances by interpolating between existing samples and their  $k$  nearest neighbors in the feature space. This process enriches minority class representation without duplication, making it a valuable tool for addressing class imbalance in real-world datasets where one class is severely underrepresented. However, while oversampling enhances model performance on minority classes, it presents challenges that can be categorized into six main areas:

1. First, one of the most common problems associated with oversampling is the potential bias towards the minority class (Tarawneh et al., 2022; Drummond and Holte, 2003). When oversampling is applied, the minority class is artificially inflated by creating new synthetic samples, leading the model to become over-reliant on this class and ignore the majority class. This can result in high accuracy on training data but poor performance on real-world data, given the infrequency of the minority class.

2. Oversampling can also lead to inconsistencies

in data types, as synthetic data points may generate values that are outside the typical range of the variable or in a different format. For example, if the original data only contains whole numbers for age, oversampling may generate decimal numbers that are not present in real-world data.

3. Synthetic samples created through oversampling are assumed to belong to the minority class, but this may not be true. It may also produce mislabeled samples belonging to the majority class, and also "noise" samples that are absurd and do not correspond to any class or reality, such as a patient aged 3 and weighing 100kg.

4. The distribution of the data may also be altered by synthetic samples. For example, if the minority class includes 50% of children but the synthesized data includes only 20% then the distribution is not the same.

5. Oversampling can reduce the diversity of the dataset by creating synthetic samples that are very similar to existing samples. This can result in overfitting and negatively impact the model's ability to generalize to new data. The oversampled dataset may not accurately reflect the true diversity of the problem.

It's important to carefully consider the impact of oversampling on the distribution and diversity of the dataset to ensure that the resulting model accurately reflects the true nature of the problem.

6. Oversampling can increase the computational cost of training a model, as it requires generating additional data points for the minority class (Chen et al., 2004; Rodríguez-Torres et al., 2022). When working with large datasets, where generating synthetic data can be time-consuming and resource-intensive.

Additionally, the more imbalanced the dataset is the less the oversampled dataset accurately reflects the true nature of the problem (He and Garcia, 2009). As explained above, the oversampling algorithm will adjust the class distribution of a data set. So the more imbalanced the dataset the more data will be a need to adjust the class distribution as a result more synthetic data the oversampled dataset contains. This can be particularly challenging when working with anomaly detection datasets since they tend to have highly imbalanced class distributions, as the occurrence of rare events or conditions is infrequent compared to the overall population. Medical and fraud detection datasets are common examples of such highly imbalanced datasets where detecting anomalies is critical, but these anomalies are rare in occurrence (Chandola et al., 2009).

Medical datasets, in particular, pose significant challenges for oversampling techniques. These datasets are often the most imbalanced because cer-

tain diseases or conditions may be rare in comparison to the overall population. For instance, a specific disease might affect only a small percentage of people, while the majority are healthy. Consequently, the dataset will have a highly imbalanced class distribution, with the minority class being the medical condition of interest (Longadge and Dongre, 2013).

Moreover, the expense and complexity of medical data collection can contribute to class imbalance. Collecting medical data often involves costly and time-intensive procedures, like medical tests or imaging, which can be difficult to perform on a large and diverse population. Consequently, the data collected may be biased towards certain groups or demographics, leading to imbalanced class distributions.

In the next section, we will illustrate some problems encountered with oversampling over a medical example.

### 3 EXPERIMENTS

#### 3.1 Description of the Initial Machine Learning Task

Our initial goal was to predict hemorrhage risk using patient medical prescriptions from the MIMIC database, a comprehensive resource of de-identified electronic health records for over 40,000 ICU patients in the United States. The database includes clinical data like demographics, diagnoses, laboratory results, medication details, and vital signs (Johnson et al., 2021).

Specifically, we sought to identify patients at high risk of hemorrhage due to specific medications, doses, and medical histories. High-risk individuals typically have a prior history of hemorrhage, which is a critical concern, as certain medications, dosages, and individual medical backgrounds can increase the likelihood of life-threatening hemorrhagic events. Common medications known to heighten hemorrhage risk include anticoagulants like warfarin, dabigatran, and apixaban, as well as antiplatelet agents like aspirin and clopidogrel. Other medications, such as non-steroidal anti-inflammatory drugs (NSAIDs) and selective serotonin reuptake inhibitors (SSRIs), may also increase the risk of hemorrhage, especially when taken in high doses or in combination with other medications (Hamrick and Nykamp, 2015).

We define the machine learning classification problem as follows:

#### **Predicting hemorrhage risk**

**Input:** Medical patient prescription history, hospital patient admission history. **Output:** Patient has a hemorrhage risk or not ?

To label the data, we first needed to define how to extract information on medication-induced hemorrhage. We achieved this by examining the patient hospital admission record, which contains the reason for admission coded using the International Classification of Diseases (ICD) system. This system is a standardized medical classification system used for coding and classifying medical procedures, symptoms, and diagnoses (World Health Organization, 2016). By analyzing the International Classification of Diseases system, we were able to define a list of ICD codes that represent medication-induced hemorrhage.

The input data included hospital admission records and current prescription details, with medications coded using the US-specific National Drug Code (NDC). However, NDC is specific to the US and is too specific, since distinct codes exist for the various dosages, forms, and presentations of a drug (U.S. Food and Drug Administration, nd). Thus, we used the Anatomical Therapeutic Chemical (ATC) classification system, which organizes medications based on their therapeutic properties and anatomical site of action (WHO Collaborating Centre for Drug Statistics Methodology, 2013). To address this difference, we mapped the NDC code to its corresponding ATC code. Some medications have multiple ATC codes; in this case, we considered all of them.

Finally, we coded patient's medication using one hot encoding. It's a process used in machine learning to convert categorical data into a numerical representation that can be used by machine learning algorithms. It involves creating a binary vector that has one value for each possible drug, the value being 1 if the drug is present and 0 otherwise. For example, if there are three medications -  $M_1$ ,  $M_2$ , and  $M_3$  - each medication or drug order would be represented by a binary vector of length three. The medication  $M_1$  would be represented by the vector  $[1,0,0]$ , the medication  $M_2$  would be represented by  $[0,1,0]$ , and the medication  $M_3$  would be represented by  $[0,0,1]$ . A drug order associating medication  $M_1$  and  $M_2$  would be represented by  $[1,1,0]$ . This allows algorithms to work with categorical data, which can be useful in many applications such as text classification.

The resulting dataset was highly imbalanced with a minority class representing only 3.47 % of patients who have a hemorrhage risk. The imbalanced nature of the dataset can pose a significant challenge for the model in accurately predicting the minority class. This is because the model may become biased

toward the majority class, which resulted in poor performance when predicting the minority class. To address this issue, we employed oversampling as a common technique to balance the dataset.

While we have presented results based on a single dataset for clarity, it is essential to note that our approach has been tested on multiple datasets (Boudegz-dame et al., 2024).

### 3.2 New Problem Encountered with Oversampling

After oversampling, we observed significant improvement in the model's performance on both the training and validation which was oversampled but performed poorly on the original data in terms of performance metrics. Moreover, predicting the risk of hemorrhage is a challenging task as it occurs infrequently, and it is difficult to predict if a prescription will result in a hemorrhage. However, we obtained an f1 score of 90% in predicting hemorrhage on training which seemed too optimistic. To investigate this issue, we conducted an analysis of the model's predictions to determine if it was still addressing the initial problem.

We formulate the hypothesis that the model was learning to predict whether a sample was synthetic, instead of predicting whether it belongs to the minority class which, indeed, is almost the same, since a large majority of the samples belonging to the minority class are synthetic.

### 3.3 A Method for Measuring the Detectability of Synthetic Data

To test this hypothesis, we defined a new machine learning problem to detect synthetic data. We generated a number of synthetic samples equal to the number of samples in the minority class using oversampling, we removed samples from the majority class, and we labeled the samples as either synthetic (1) or original (0). We applied this approach to different oversampling methods. It aims at determining the ease with which synthetic data generated by these methods could be detected, with lower-quality data being more easily detected. Our refined dataset was used to address the following problem:

#### *Detecting synthetic data*

**Input:** *Minority class VS synthetic data produced by oversampling. Output:* *Is the instance synthetic or original?*

In our analysis, we considered a range of evaluation metrics to assess the model's performance:

1. **Precision, Recall, and F1 Score:** Precision mea-

sures correct positive predictions out of all predicted positives, while recall measures correct positives out of all actual positives. F1 score, the harmonic mean of precision and recall, assesses model performance, especially with highly imbalanced data (He and Garcia, 2009; Powers, 2011).

2. **Area Under the Precision-Recall Curve (AUPRC):** A single score capturing the trade-off between precision and recall, especially valuable for imbalanced data as it focuses on the positive class and can provide a more informative evaluation than accuracy or ROC AUC (Davis and Goadrich, 2006).
3. **Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC):** ROC depicts true positive rates versus false positive rates at various decision thresholds, while AUC condenses this curve into a single performance score. These metrics are valuable for comparing models with varying thresholds (He and Garcia, 2009; Powers, 2011; Fawcett, 2006).
4. **Confusion Matrix:** Providing detailed insights into true positives, true negatives, false positives, and false negatives. This helps identify correct and incorrect classifications for each class.
5. **Cohen's kappa:** measures inter-rater agreement between the original and oversampled datasets. It can be useful for evaluating how well the synthetic data captures the true nature of the problem (McHugh, 2012).

When learning to predict whether samples are synthetic, we obtain performance metric values that indicate the success of the learning process. These metric values serve as a measure that quantifies the problem we discovered.

In this second learning task, it is crucial to use the same machine learning technique as in the initial learning process. This consistency ensures a fair test to determine whether this technique can effectively discern the synthetic nature of the samples, as using a different technique may behave differently. Therefore, we have opted for a neural network.

For the current implementation, we used a neural network with two hidden layers containing 30 and 20 neurons respectively. To prevent the issue of "dead" neurons in deep networks, we opted for the *LeakyReLU* activation function, which has been shown to perform well in similar applications (He et al., 2016). The output layer was designed with a *sigmoid* activation function, commonly used in binary classification problems.

To optimize training, we employed *ReduceLROn-Plateau* learning rate scheduling. This technique al-

lowed us to dynamically adjust the learning rate of the optimizer during training, based on a monitored metric such as validation loss. By doing so, we were able to help the model escape plateaus and continue to improve, even as it approached convergence. Our model was trained over 100 epochs, which was sufficient to ensure full learning and convergence of the model.

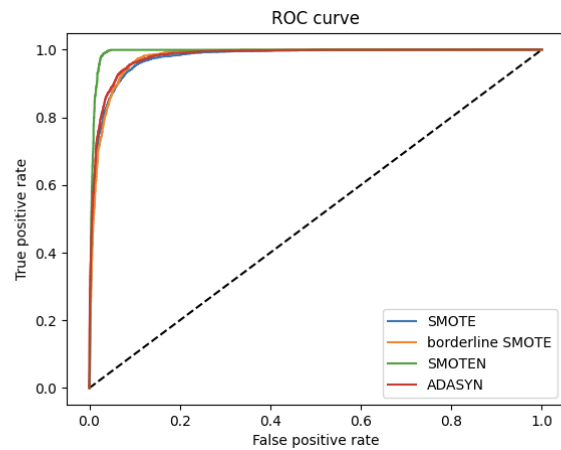
### 3.4 Results and Analysis

The results in the table 1 demonstrate that the neural network performed exceptionally well in predicting synthetic samples in terms of the evaluation metrics across all four oversampling techniques. Notably, both precision and recall are consistently high and similar, indicating an absence of bias toward the majority class. This suggests that the model effectively identifies both synthetic and original data instances, which is particularly noteworthy given that the model was trained on data where the synthetic class represents the majority of the samples (about 95%).

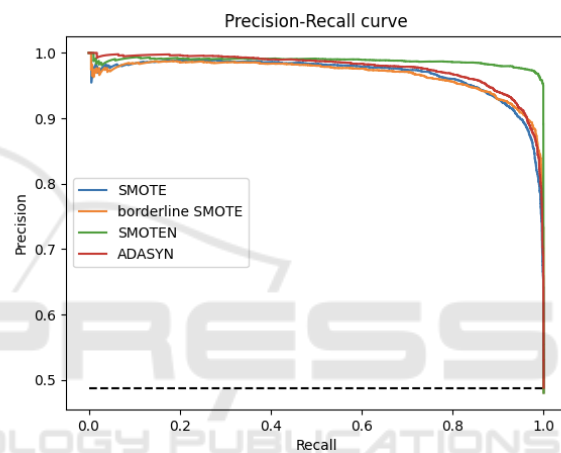
The highest score was achieved for SMOTEN in terms of f1 score, recall, precision, accuracy, cohen kappa, and AUC among all oversampling techniques. The Borderline SMOTE algorithm also leads to high scores in all evaluation metrics except for AUC. Therefore, we can easily predict whether a sample is synthetic or not. This prediction is much easier than predicting hemorrhage risk. Thus it confirms our hypothesis: the initial model was in fact predicting the synthetic nature of data instead of hemorrhage risk.

As explained above, the ROC curve and Precision-Recall curve provide important information about the performance of a binary classification model. Therefore, we plotted both curves to obtain a more comprehensive evaluation of the model's performance. Figure 1a summarise the ROC curve for the four oversampling algorithms. It indicates that the model has high accuracy in distinguishing between positive and negative samples. In fact, an AUC of 0.5 suggests a random classification, while an AUC of 1 suggests a perfect classification. The AUC values for SMOTE, borderlineSMOTE and ADASYN are 0.97, indicating that the model's performance is very close to perfect, with only a small number of false positives and false negatives. Furthermore, we observed that the oversampled data generated by SMOTEN on our data were the easiest to detect, as confirmed by Figure 1b, which summarises the Recall-Precision curve.

Therefore, the table 1 and the figures 1a and 1b suggest that oversampling techniques can be easily detected to a great extent. However, the choice of the oversampling algorithm should depend on the specific characteristics of the dataset and the evaluation met-



(a) ROC Curves for Oversampling Techniques.



(b) Precision-Recall Curves for Oversampling Techniques.

Figure 1: Performance Evaluation of Oversampling Techniques: ROC and Precision-Recall Curves.

rics of interest.

While it's known that oversampling algorithm does not behave the same on different dataset, The testing results on medical data including medical prescription, which is a highly imbalanced dataset and strongly indicates that oversampling will not be a considerable technique for balancing our data. Further analysis and experimentation may be necessary to determine the most effective approach for balancing the medical prescription dataset in question.

### 3.5 Understanding Why Synthetic Data Are Easily Detected

Oversampled medication prescriptions may not accurately represent real-world data, as they are easily detectable by machine learning algorithms. To gain a better understanding of this issue, we have formulated the following hypotheses:

Table 1: Performance comparison of oversampling algorithms on synthetic data classification.

|                         | F1 Score | Recall | Precision | Accuracy | Cohen Kappa | AUC  |
|-------------------------|----------|--------|-----------|----------|-------------|------|
| <b>SMOTE</b>            | 0.92     | 0.94   | 0.90      | 0.92     | 0.84        | 0.97 |
| <b>Borderline SMOTE</b> | 0.93     | 0.96   | 0.91      | 0.93     | 0.86        | 0.97 |
| <b>SMOTEN</b>           | 0.97     | 0.99   | 0.96      | 0.97     | 0.95        | 0.99 |
| <b>ADASYN</b>           | 0.92     | 0.93   | 0.91      | 0.92     | 0.84        | 0.97 |

**Hypothesis 1: Over or under representation of drugs in synthetic data. (Problem #4 in section 3)** Medical prescriptions typically contain a limited number of medications. However, synthetic data generated may contain a smaller or larger number of medications, resulting in an under and over representation of drugs respectfully, which could lead to discrepancies between the synthetic and real-world data.

The following table 2 shows that all four oversampling methods (SMOTE, Borderline SMOTE, SMOTEN, and ADASYN) have resulted in a decrease in the mean number of ATC codes for medication in the oversampled data compared to the original data. This indicates an under-representation of medication in the oversampled samples.

Table 2: Drug distribution in original and synthetic data.

| Dataset                 | Mean number of ATC codes |
|-------------------------|--------------------------|
| <b>Original</b>         | 34.78                    |
| <b>SMOTE</b>            | 20.11                    |
| <b>Borderline SMOTE</b> | 21.13                    |
| <b>SMOTEN</b>           | 20.76                    |
| <b>ADASYN</b>           | 19.49                    |

**Hypothesis 2: Changes in the nature of data. (Problem #2)** SMOTE can introduce small perturbations to feature values in order to create synthetic samples, which may result in non-integer or floating-point values for discrete features (Blagus and Lusa, 2013). For example, drugs are represented as discrete values of 0 or 1, indicating the presence or absence of the drug in a prescription. However, synthetic data generated for the purpose of analysis may contain drugs with continuous values, which may lead to inaccuracies in the results.

After further investigation, we found that the application of SMOTE, Borderline SMOTE, SMOTEN, and ADASYN did not result in any significant changes to the nature of the oversampled data. All four oversampling methods applied to our data did not alter the nature of the data.

**Hypothesis 3: Inconsistencies in ATC codes. (Problem #3)** Some drugs such as aspirin have several ATC codes, and we associated them with all of their

corresponding codes in the original data. However, in the synthetic samples, such a drug may be associated with only one of its codes. For instance, an aspirin prescription might be coded as a platelet aggregation inhibitor but not as an analgesic in the synthetic samples.

**Hypothesis 4: Inconsistencies in drug associations. (Problem #3)** Synthetic prescriptions generated may include inconsistent drug associations. For instance, drugs such as ramipril and enalapril, which are both angiotensin-converting-enzyme inhibitors and have the same effects, thus they are never associated together. However, such inconsistencies may occur in the synthetic samples.

## 4 DISCUSSION

In this paper, we address a common problem associated with oversampling: the risk of the machine learning model learning to detect the synthetic nature of oversampled data rather than the original underlying patterns. We propose a method to identify and quantify this problem, focusing on the extent to which synthetic data can be detected.

In the literature many studies have explored the problems associated with oversampling and SMOTE, however, to the best of our knowledge, none of them neither mentioned the learning of the synthetic nature of data nor proposed a method for quantifying it.

Tarawneh and al. (Tarawneh et al., 2022) provide a comprehensive review of class imbalance mitigation in machine learning, focusing on oversampling. They highlight issues like overfitting, higher computational costs, and reduced generalization performance. The article also emphasizes the risks of model bias and decreased generalization when oversampled data is tested on original database, along with the significant computational overhead of creating and storing synthetic samples. The authors suggest alternative approaches like cost-sensitive learning and anomaly detection as more effective solutions to tackle class imbalance.

The work of R. Buda and al. (Buda et al., 2018) investigates the impact of class imbalance on CNN

performance for image classification tasks and evaluates various strategies, including oversampling. They caution that oversampling alone may not suffice for addressing class imbalance in CNNs due to the risk of overfitting, where models memorize training data and perform poorly on new data. Furthermore, oversampling can generate unrealistic and redundant samples, inefficiently utilizing computational resources.

Several studies propose modifications to the oversampling technique to mitigate these issues. Rodríguez-Torres and al. (Rodríguez-Torres et al., 2022) introduce Large-scale Random Oversampling (LRO) to address class imbalance in large datasets. Comparisons with other oversampling methods, such as SMOTE and Borderline-SMOTE, show that LRO achieves higher accuracy and F1-score while being computationally efficient. The study highlights SMOTE's limitations, including sample diversity issues and sensitivity to noise.

Overall, the literature highlights the potential limitations and challenges of oversampling and SMOTE in addressing imbalanced data in machine learning, and suggests alternative approaches and modifications to address these issues. The presented articles cover various aspects of oversampling and SMOTE problems, including overfitting, performance evaluation, large dataset handling, multi-class imbalance, noise handling, and synthetic oversampling.

## 5 CONCLUSION AND PERSPECTIVES

In conclusion, oversampling is a valuable tool for improving machine learning model performance on imbalanced datasets. However, our research highlights the potential issues introduced by oversampling algorithms, particularly in the quality of synthetic minority class data, which can lead to models learning to predict noise rather than underlying patterns. To address these concerns, we have proposed a novel evaluation method that assesses and quantifies both the effectiveness of oversampling techniques and their potential to introduce detectable noise. By evaluating a model's ability to differentiate synthetic data from real data, we can identify potentially problematic oversampling methods and select the most suitable ones for specific datasets, ultimately enhancing model accuracy and generalizability (Boudegzdame et al., 2024). This approach also aids in determining the suitability of oversampling for dataset balancing.

The perspectives of this study are: 1) delimiting the exact perimeter of the problem we discovered, in particular testing other similar existing oversampling

techniques, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), 2) improving the measure we proposed for quantifying the detectability of synthetic data, for instance for multi-class and/or multi-label classification, and 3) designing new methods of oversampling that are resilient to this problem.

## ACKNOWLEDGEMENTS

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

## REFERENCES

- Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–30.
- Blagus, R. and Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:106.
- Boudegzdame, N., Sedki, K., Tspora, R., and Lamy, J.-B. (2024). An approach for improving oversampling by filtering out unrealistic synthetic data. *ICAART 2024*.
- Buda, R., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.
- Bunghumpornpat, 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.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling 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. Technical Report 110, University of California, Berkeley.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240.
- Drummond, C. and Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

- 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.
- Hamrick, J. W. and Nykamp, D. (2015). Drug-induced bleeding. *US Pharmacist*, 40(12):HS17–HS21.
- 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 *IEEE International Joint Conference on Neural Networks*, pages 1322–1328.
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501.
- Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng, S. (2021). Mimic-iv (version 1.0). PhysioNet.
- Longadge, R. and Dongre, S. (2013). Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.
- Powers, D. M. W. (2011). Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.
- 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.
- U.S. Food and Drug Administration (n.d.). National drug code (ndc) directory.
- WHO Collaborating Centre for Drug Statistics Methodology (2013). Guidelines for atc classification and ddd assignment 2013. Technical report, World Health Organization, Oslo, Norway.
- World Health Organization (2016). *International Classification of Diseases, 11th Revision (ICD-11)*. Geneva.