SMOTE in Predictive Modeling: A Comprehensive Evaluation of Synthetic Oversampling for Class Imbalance

Sandeep Yadav

Silicon Valley Bank, Tempe, USA
sandeep.yadav@asu.edu
ORCID: 0009-0009-2846-0467

Abstract

Class imbalance is a pervasive challenge in predictive modeling, where minority class instances are significantly underrepresented, leading to biased models and suboptimal performance. Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely used solutions to address this issue by generating synthetic samples for the minority class. This study provides a comprehensive evaluation of SMOTE and its variants in handling class imbalance across diverse datasets and model types. We assess SMOTE's impact on predictive performance, model generalizability, and stability under different imbalance ratios. Key SMOTE variations, including Borderline-SMOTE, SMOTE-ENN, and SMOTE-Tomek, are applied and analyzed across metrics such as precision, recall, F1-score, and area under the precision-recall curve (AUC-PR). Experimental results reveal that while SMOTE enhances model performance by providing additional decision boundaries for the minority class, certain variants, such as SMOTE-ENN, excel in high-dimensional spaces by removing noisy samples post-oversampling. However, SMOTE can introduce synthetic noise when applied without consideration of data density and class boundaries. Overall, findings highlight the effectiveness of SMOTE and its variants in improving minority class prediction but underscore the importance of selecting the appropriate variant based on dataset characteristics and desired performance metrics. This study provides practical guidance for data scientists and researchers on utilizing SMOTE for imbalanced datasets, promoting robust and fair predictive models in diverse real-world applications.

Key Words: **Class Imbalance, Noise Reduction, Minority class, Undersampling, Oversampling, SMOTE (Synthetic Minority Over-sampling Technique), Data Augmentation Techniques, Precision-Recall Curve (AUC-PR), Model Generalizability**

1. INTRODUCTION

Class imbalance is a critical and frequent challenge in predictive modeling, where certain classes—often the ones with the most significance, such as rare diseases or fraudulent transactions—are vastly underrepresented in comparison to other classes. This imbalance skews predictive performance, leading most machine learning algorithms to favor the majority class. As a result, predictions on the minority class can suffer from low recall and precision, undermining model effectiveness in applications where minority classes are crucial. Tackling class imbalance has therefore become essential in fields ranging from healthcare and finance to text analytics and anomaly detection.

One of the most popular methods for handling imbalanced datasets is the Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. in 2002. SMOTE addresses class imbalance by generating synthetic samples for the minority class rather than duplicating existing instances. It does so by interpolating

between existing minority samples, creating new instances that can help the classifier better define decision boundaries for the minority class. This approach has led SMOTE to become a widely used preprocessing step, effectively increasing the representation of the minority class without risking overfitting due to sample repetition.

However, the utility of SMOTE varies depending on dataset characteristics and model complexity. To address this, several SMOTE variants have been developed, each tailored to different dataset conditions. Borderline-SMOTE focuses on generating samples closer to the decision boundary between classes, aiming to improve performance in cases with highly overlapping classes. SMOTE-ENN and SMOTE-Tomek combine SMOTE with undersampling techniques, removing noisy or ambiguous samples after oversampling. These hybrid methods have shown promise in reducing class overlap and enhancing model robustness by balancing sample augmentation and noise reduction. Yet, the choice of the most suitable SMOTE variant remains dataset-dependent, as each variant has unique effects on model performance.

The effectiveness of SMOTE and its variants across diverse contexts has not been systematically evaluated, despite its widespread use. While SMOTE is often applied as a universal solution to class imbalance, certain variants may outperform others depending on factors like imbalance ratio, feature dimensionality, and noise levels. For example, SMOTE can introduce synthetic noise if the dataset is highly sparse or if synthetic samples extend beyond the minority class boundary, leading to reduced precision. Conversely, techniques like SMOTE-Tomek may improve recall in high-dimensional spaces by minimizing class overlap. Understanding the nuanced effects of SMOTE and its variants on predictive modeling is essential for practitioners who wish to select the most effective approach based on their specific dataset and modeling goals.

This paper aims to provide a comprehensive evaluation of SMOTE and its variants in addressing class imbalance across multiple datasets and predictive models. We will analyze the impact of SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-Tomek on various machine learning models, focusing on metrics such as precision, recall, F1-score, and area under the precision-recall curve (AUC-PR). By testing these techniques under different imbalance ratios and model complexities, this study seeks to identify best practices for applying SMOTE in real-world scenarios.

2. LITERATURE REVIEW

Class imbalance in machine learning has received significant research attention due to its impact on model performance, particularly in applications where minority class prediction is crucial. Various methods have been developed to address class imbalance, including cost-sensitive learning, algorithm modification, and data-level approaches. Among data-level approaches, the Synthetic Minority Over-sampling Technique (SMOTE) and its variants have become foundational techniques for generating synthetic data points in the minority class. This section provides an overview of SMOTE and its advanced versions, examining their effectiveness, underlying algorithms, and use cases, and concludes with a visual illustration of the SMOTE architecture.

2.1 Synthetic Minority Over-sampling Technique (SMOTE)

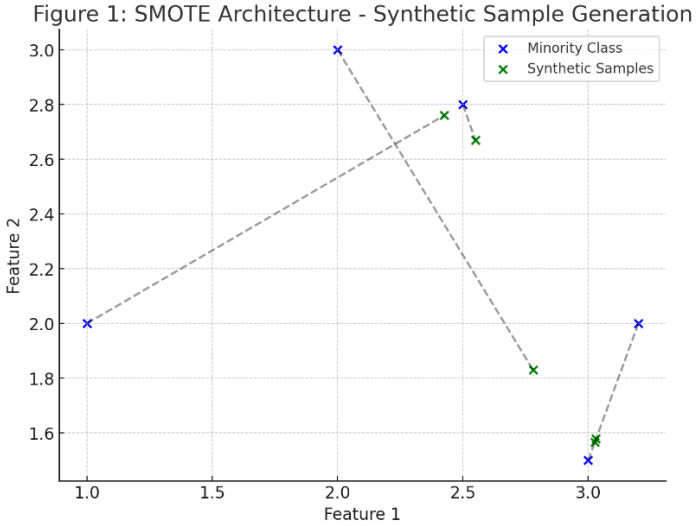
Introduced by Chawla et al. (2002), SMOTE addresses class imbalance by generating synthetic samples through interpolation between existing minority class samples. Unlike traditional oversampling, which duplicates minority instances and risks overfitting, SMOTE introduces diversity by creating new samples along the line segments between randomly selected pairs of minority class instances. Mathematically, for a

given minority instance x_i and one of its *k*-nearest neighbors$x_{neighbor}$, SMOTE generates a synthetic instancex_{new} as follows:

$$x_{new} = x_i + \delta.(x_{neighbor} - x_i)$$

where δ is a random number between 0 and 1. This formula ensures that the synthetic points are linearly interpolated between existing minority points, thus enhancing the minority class distribution. Figure 1 below illustrates SMOTE's architecture, where synthetic samples (green points) are generated along the line segments connecting minority samples (blue points).

Figure 1: SMOTE Architecture - Synthetic Sample Generation



In this diagram, SMOTE's synthetic samples fill the feature space around minority points, providing better class representation without redundant duplication. However, while SMOTE improves recall for the minority class by expanding its distribution, it may inadvertently create synthetic samples in regions that overlap with the majority class, leading to noise.

2.2 Variants of SMOTE

To address the limitations of standard SMOTE, several variants have been developed, each with unique adjustments to enhance its effectiveness for different types of data. Key SMOTE variants include Borderline-SMOTE, SMOTE-Tomek, and SMOTE-ENN, each of which optimizes the sample generation and noise reduction process.

2.2.1 Borderline-SMOTE

Borderline-SMOTE, proposed by Han et al. (2005), addresses one of SMOTE's key limitations: the generation of synthetic samples in both safe (low-risk) and risky regions near the majority class. In many cases, samples near the decision boundary between classes are more critical to model performance. Borderline-SMOTE focuses on creating synthetic samples only in these "borderline" regions, where the minority class instances are likely to be misclassified. The algorithm identifies minority instances near the majority class by examining the *k*-nearest neighbors and generating synthetic samples closer to the class boundary.

The procedure can be described as follows:

- Identify minority instances near the decision boundary (those with a higher proportion of majority class neighbors).

- Generate synthetic samples along the line segment between each borderline instance and its selected minority neighbors.

Borderline-SMOTE enhances model performance in imbalanced data scenarios with high overlap between classes. However, it may still generate some overlap between classes if the boundary is not well-defined.

2.2.2 SMOTE-Tomek Links

SMOTE-Tomek, a hybrid approach combining SMOTE with Tomek Links, was designed to further improve class separation. Tomek Links are pairs of samples, each from different classes, that are each other's nearest neighbors. When such pairs exist, it implies that these instances lie near the decision boundary and may contribute to class overlap or noise. In SMOTE-Tomek, synthetic samples are first generated using SMOTE, and then Tomek Links are applied to remove overlapping samples.

This approach reduces noise and enhances class separation. The SMOTE-Tomek method can be outlined as follows:

- Apply SMOTE to generate synthetic minority samples.
- Identify Tomek Links between majority and minority instances.
- Remove Tomek Links, thereby improving the distinction between classes and reducing noisy samples.

By eliminating potentially overlapping instances, SMOTE-Tomek achieves better separation between classes, leading to improved classifier performance.

2.2.3 SMOTE-ENN (Edited Nearest Neighbors)

SMOTE-ENN combines SMOTE with Edited Nearest Neighbors (ENN) to further enhance class separation and reduce noise. ENN removes any instance (either majority or minority) that is misclassified by its k-nearest neighbors, which reduces the number of overlapping samples and isolates the minority class. The SMOTE-ENN method is particularly effective in high-dimensional spaces where class overlap is prevalent.

The process of SMOTE-ENN is as follows:

- Generate synthetic samples using SMOTE to augment the minority class.
- Apply ENN to remove noisy or misclassified samples from the dataset.

SMOTE-ENN is advantageous in datasets with substantial class overlap and high feature dimensionality, where removing noisy samples is critical for accurate classification.

2.3 Comparison of SMOTE Variants

Each SMOTE variant has specific advantages and limitations based on the dataset characteristics:

- Standard SMOTE is effective for moderately imbalanced datasets but may introduce synthetic noise in regions near the majority class.
- Borderline-SMOTE is ideal for datasets with clear decision boundaries, as it focuses on samples near the boundary and avoids oversampling in safe zones.
- SMOTE-Tomek effectively reduces class overlap by removing Tomek Links, enhancing minority class separation while retaining critical samples.

- SMOTE-ENN removes noise by combining oversampling with ENN, making it suitable for high-dimensional data with significant overlap.

These variants enable practitioners to tailor synthetic oversampling techniques to specific data characteristics, ultimately improving model performance on imbalanced datasets.

III. EXPERIMENT SETUP

The following methodology outlines the experimental design for comparing SMOTE and its key variants in handling class imbalance across multiple datasets and machine learning models. This section includes data selection, SMOTE technique applications, model training, evaluation metrics, and the experimental procedure.

Here we take the Credit Card Fraud Detection Dataset. It has less than 0.15% fraudulent transactions, which makes it an extreme class imbalance dataset. Now we will apply the different SMOTE variants as discussed above in the paper to solve the imbalance problem. We use the *imblearn.over_sampling* library of python to try different SMOTE techniques as follows:

a) SMOTE:

```
from imblearn.over_sampling import SMOTE

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using SMOTE
smt = SMOTE()
#X_train, y_train = smt.fit_resample(X_train, y_train)
X_train_sm, y_train_sm = smt.fit_resample(X_train, y_train)

counter = Counter(y_train_sm)
print('After',counter)
```

```
Before Counter({0: 18497, 1: 4208})
After Counter({0: 18497, 1: 18497})
```

b) Borderline SMOTE

```
from imblearn.over_sampling import BorderlineSMOTE

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using ADASYN
blsmt = BorderlineSMOTE(sampling_strategy='auto', random_state=42)
X_train_blsmt, y_train_blsmt = blsmt.fit_resample(X_train, y_train)

counter = Counter(y_train_blsmt)
print('After',counter)
```

```
Before Counter({0: 18497, 1: 4208})
After Counter({0: 18497, 1: 18497})
```

c) SMOTE + Tomek

```
from imblearn.combine import SMOTETomek

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using SMOTE + Tomek
smtom = SMOTETomek(random_state=139)
X_train_smtom, y_train_smtom = smtom.fit_resample(X_train, y_train)

counter = Counter(y_train_smtom)
print('After',counter)
```

```
Before Counter({0: 18497, 1: 4208})
After Counter({0: 18090, 1: 18090})
```

d) SMOTE + ENN

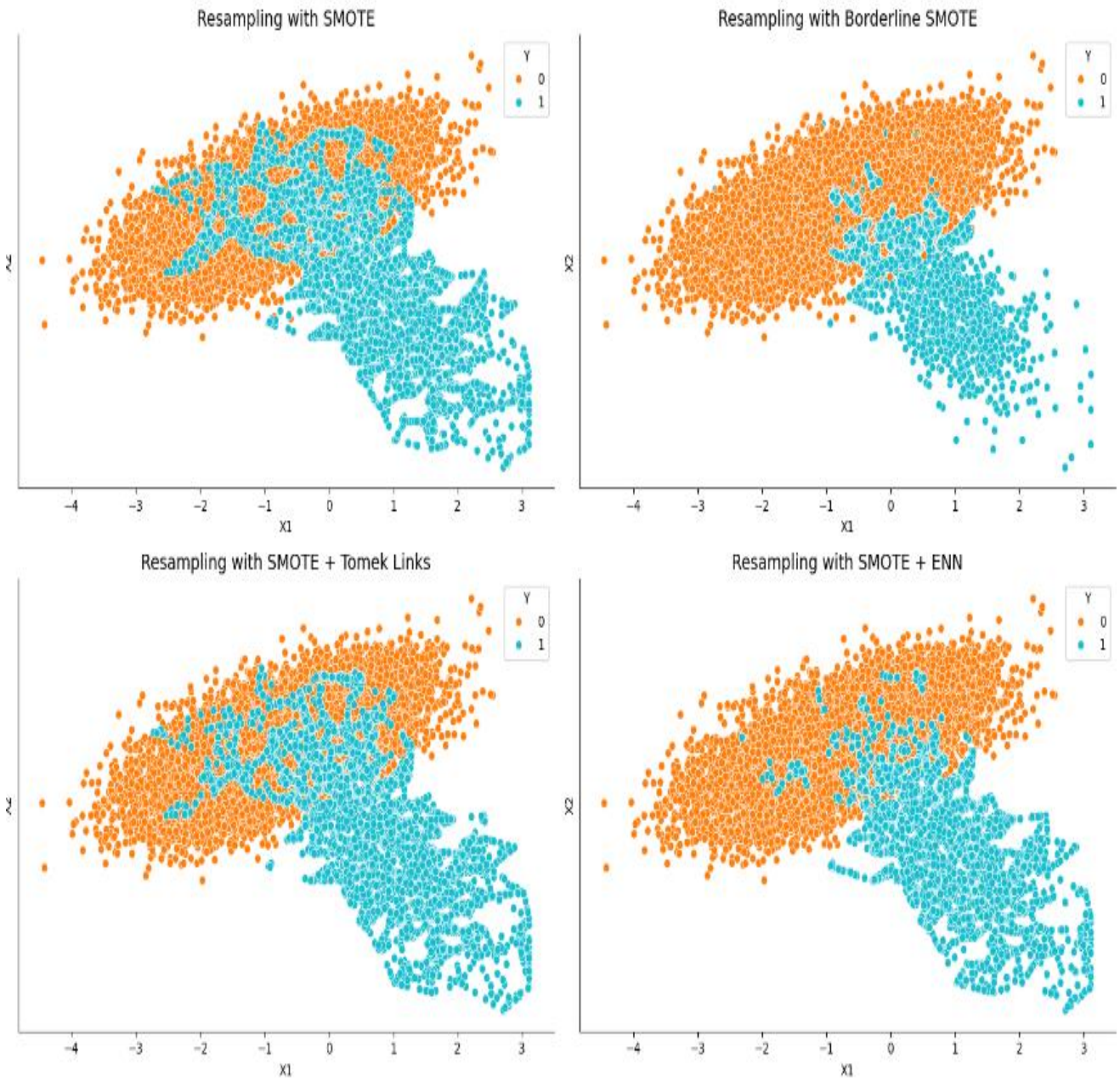
```
from imblearn.combine import SMOTEENN

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using SMOTE + ENN
smenn = SMOTEENN()
X_train_smenn, y_train_smenn = smenn.fit_resample(X_train, y_train)

counter = Counter(y_train_smenn)
print('After',counter)
```

```
Before Counter({0: 18497, 1: 4208})
After Counter({1: 14813, 0: 8948})
```

The below-given Figure 2 shows how different SMOTE based resampling techniques used above will work out to deal with imbalanced data.



VI. RESULTS & EVALUATION

A comparative analysis was done on the fraud dataset using different models for classification: Decision Tree, Random Forest and Logistics Regression. Here we will ignore the accuracy metric to evaluate the performance for this imbalanced dataset. We are more interested to know which will be the fraudulent transactions in the future data. Hence we will compare the performance of metrics like precision, recall, F1-score and AUC-ROC to understand the performance of the classifiers.

From the below Figure 3 we can see that on the actual imbalance dataset, all the 3 classifier models were not able to generalize much on the minority class versus the majority class. Resulting in classifying correctly most of the negative class samples. Because of this there were less False Positives compared to more False Negatives. After using the different oversampling techniques, we can clearly see quite a bit of increase in the Recall on the test data.

Model	Resample	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	Actual	0.745614	0.080798	0.145798	0.773214
Logistic Regression	SMOTE	0.448557	0.679658	0.540438	0.778493
Logistic Regression	Borderline SMOTE	0.433779	0.678707	0.529281	0.777567
Logistic Regression	SMOTE + Tomek	0.446173	0.681559	0.539300	0.778341
Logistic Regression	SMOTE + ENN	0.450311	0.689163	0.544703	0.782601
Decision Tree	Actual	0.669591	0.435361	0.527650	0.802032
Decision Tree	SMOTE	0.402319	0.659696	0.499820	0.782408
Decision Tree	Borderline SMOTE	0.396689	0.660646	0.495720	0.788294
Decision Tree	SMOTE + Tomek	0.398319	0.675856	0.501234	0.785032
Decision Tree	SMOTE + ENN	0.325852	0.772814	0.458416	0.781943
Random Forest	Actual	0.741883	0.434411	0.547962	0.829965
Random Forest	SMOTE	0.440303	0.662548	0.529032	0.805693
Random Forest	Borderline SMOTE	0.442643	0.674905	0.534639	0.808772
Random Forest	SMOTE + Tomek	0.454253	0.670152	0.541475	0.809254
Random Forest	SMOTE + ENN	0.366605	0.757605	0.494110	0.800162

To understand this better a bar chart is plotted showing the comparative study of results from all the 3 classifiers in figure 4 below.

V. CONCLUSION

The results reveal that SMOTE-based techniques can significantly enhance model performance on minority classes, particularly in highly imbalanced scenarios where traditional training methods often fail. However, the effectiveness of each SMOTE variant varies depending on dataset characteristics and model type, underscoring the importance of selecting the appropriate technique based on the specific requirements of a predictive modeling task.

Standard SMOTE demonstrates consistent improvements in recall and F1-score across simpler models like logistic regression, making it a reliable choice for general applications. Borderline-SMOTE, with its focus on decision boundaries, enhances precision and recall in datasets with high class overlap, particularly when misclassification costs are critical. The hybrid techniques, SMOTE-Tomek and SMOTE-ENN, stand out for their noise reduction capabilities. SMOTE-ENN achieves the highest overall performance by combining

synthetic sample generation with noise removal, which is especially beneficial in high-dimensional spaces and for complex models like Random Forest.

This analysis highlights that while SMOTE and its variants are powerful tools for managing imbalanced data, no single technique is universally superior. Instead, each technique's success depends on factors like the imbalance ratio, feature dimensionality, and model complexity. Practitioners should carefully consider these aspects when choosing a SMOTE technique, balancing synthetic oversampling with potential noise reduction to achieve optimal predictive performance. This study offers practical guidance for data scientists and researchers, encouraging the informed application of SMOTE to promote fair and robust predictive models in imbalanced data contexts.

REFERENCES

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357. doi:10.1613/jair.953
2. Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning. *Proceedings of the 2005 International Conference on Intelligent Computing*, 878-887. Springer, Berlin, Heidelberg. doi:10.1007/11538059_91
3. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations*, 6(1), 20-29. doi:10.1145/1007730.1007735
4. Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-sampling Technique for Handling the Class Imbalanced Problem. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 475-482. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-01307-2_43
5. Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2009). Experimental Perspectives on Learning from Imbalanced Data. *Proceedings of the 24th ACM Symposium on Applied Computing*, 782-787. doi:10.1145/1529282.1529469
6. He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. doi:10.1109/TKDE.2008.239
7. Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 39(2), 539-550. doi:10.1109/TSMCB.2008.2007853
8. Fernandez, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. *Springer International Publishing*. doi:10.1007/978-3-319-98074-4
9. Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Proceedings of the 14th International Conference on Machine Learning*, 179-186. Morgan Kaufmann Publishers Inc.
10. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Networks*, 106, 249-259. doi:10.1016/j.neunet.2018.07.011
11. Yen, S. J., & Lee, Y. S. (2009). Cluster-Based Under-Sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications*, 36(3), 5718-5727. doi:10.1016/j.eswa.2008.06.108
12. Blagus, R., & Lusa, L. (2013). SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinformatics*, 14(1), 106. doi:10.1186/1471-2105-14-106
13. Douzas, G., & Bacao, F. (2018). Geometric SMOTE: A Geometric Approach to Data Augmentation in Imbalanced Learning. *Expert Systems with Applications*, 90, 207-217. doi:10.1016/j.eswa.2017.08.04