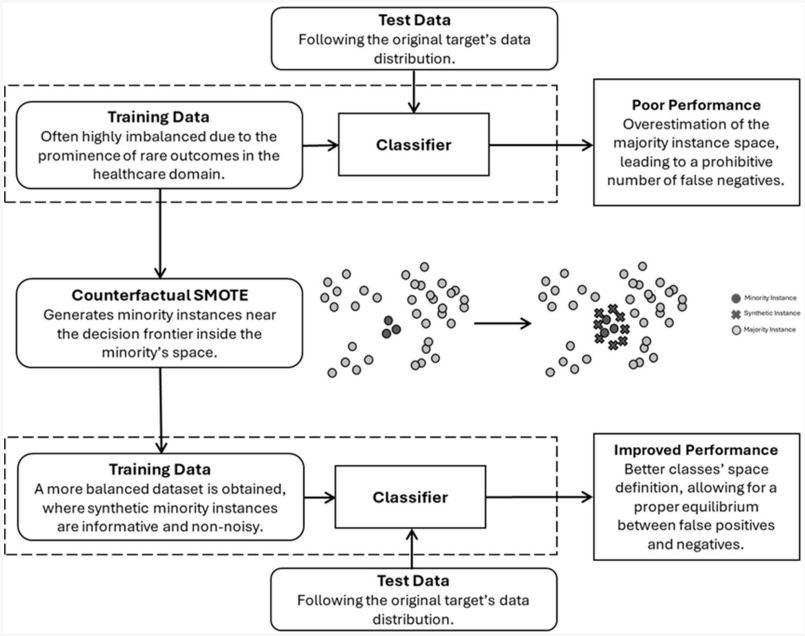
New AI meA groundbreaking machine learning technique, Counterfactual

GA, UNITED STATES, April 16, 2025 /EINPresswire.com/ -- A novel approach called Counterfactual Synthetic Minority Oversampling Technique (SMOTE) has been developed to tackle the persistent issue of imbalanced data in healthcare. Traditional models trained on imbalanced datasets often overlook rare but critical conditions, such as diseases, leading to biased predictions. By combining SMOTE with counterfactual generation, Counterfactual SMOTE creates synthetic data points near decision boundaries while minimizing noise. This method significantly enhances model performance, enabling better detection of rare conditions without overwhelming false positives. Tested across 24 healthcare datasets, Counterfactual SMOTE outperforms existing methods, offering a robust solution for improving medical diagnostics and extending to other fields.



Schematization of the application of the Counterfactual synthetic minority oversampling technique (SMOTE) to an imbalanced learning problem.

Machine learning holds great promise in healthcare, with applications ranging from early disease detection to personalized treatments. However, its effectiveness is often hindered by imbalanced data, where rare, critical outcomes such as certain diseases are vastly underrepresented compared to negative cases. As a result, traditional models tend to favor the majority class, neglecting life-threatening conditions. While techniques like Synthetic Minority Oversampling Technique (SMOTE) attempt to balance these datasets by generating synthetic minority samples, they often produce noisy or redundant data, leading to misdiagnoses or wasted resources. Addressing these shortcomings, there is a need for advanced methods that can improve model accuracy and reliability without introducing unwanted noise.

On January 25, 2025, researchers Goncalo Almeida and Fernando Bacao from NOVA Information

Management School introduced Counterfactual SMOTE, a new enhancement to the widely used SMOTE technique. Published (DOI: 10.1016/j.dsm.2025.01.006) in Data Science and Management, this new method integrates counterfactual generation to place synthetic samples strategically near decision boundaries within the "safe" minority regions. Validated on 24 highly imbalanced healthcare datasets, Counterfactual SMOTE showed a 10% average improvement in F1-score, significantly outperforming existing methods. This innovation marks a major step forward in addressing the challenges of imbalanced data, offering improved performance for medical diagnostics and beyond.

Counterfactual SMOTE improves upon traditional SMOTE by addressing two critical issues: noisy samples and near-duplicates. It generates synthetic data points as counterfactuals of majority-class instances, ensuring that these samples are placed near the decision boundary, where misclassification risks are highest. By utilizing a binary search along the line connecting majority and minority samples, guided by a k-NN classifier, the method ensures that synthetic data remains within "minority-safe" zones, thereby reducing potential noise. Key innovations include boundary-focused sampling, which uses majority-minority pairs rather than interpolating between minority samples. The method has been validated across eight benchmark models, including Borderline SMOTE and Adaptative Synthetic Sampling Method (ADASYN), showing significant improvements in reducing false negatives by 24%–34% while maintaining low false positives. Although the method incurs higher computational costs, the gains in accuracy, particularly in resource-critical fields like healthcare, justify its application. Moreover, its generalizability extends beyond healthcare, making it applicable to other domains like fraud detection and manufacturing defect analysis.

Dr. Goncalo Almeida, the study's lead author, emphasized, "Counterfactual SMOTE bridges the gap between data imbalance and actionable AI. By focusing on safe, informative samples, it ensures models don't just 'guess' majority classes but truly learn to identify rare cases. This is a paradigm shift for imbalanced learning, with life-saving implications in medical diagnostics." Dr. Almeida highlighted the method's potential to enhance the precision of AI models in healthcare, ensuring that they prioritize rare conditions without overwhelming the system with false alarms. This breakthrough represents a transformative step in the field of imbalanced data learning.

Counterfactual SMOTE's impact extends well beyond healthcare. In sectors like finance, the method could improve fraud detection by ensuring that rare fraudulent activities are accurately identified, while in telecommunications, it could predict customer churn with higher precision. In healthcare, the method enables accurate detection of rare diseases, balancing the need for precise identification with the prevention of false positives that can overwhelm healthcare systems. Open-sourcing the code further facilitates broader adoption across industries. Future developments may explore expanding the method’s capabilities to handle categorical data and multiclass applications, reinforcing Counterfactual SMOTE as a cornerstone solution for tackling data imbalance in a wide range of fields.

References

DOI
doi.org/10.1016/j.dsm.2025.01.006

Original Source URL
https://doi.org/10.1016/j.dsm.2025.01.006

Funding information
This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS).

Lucy Wang
BioDesign Research
email us here