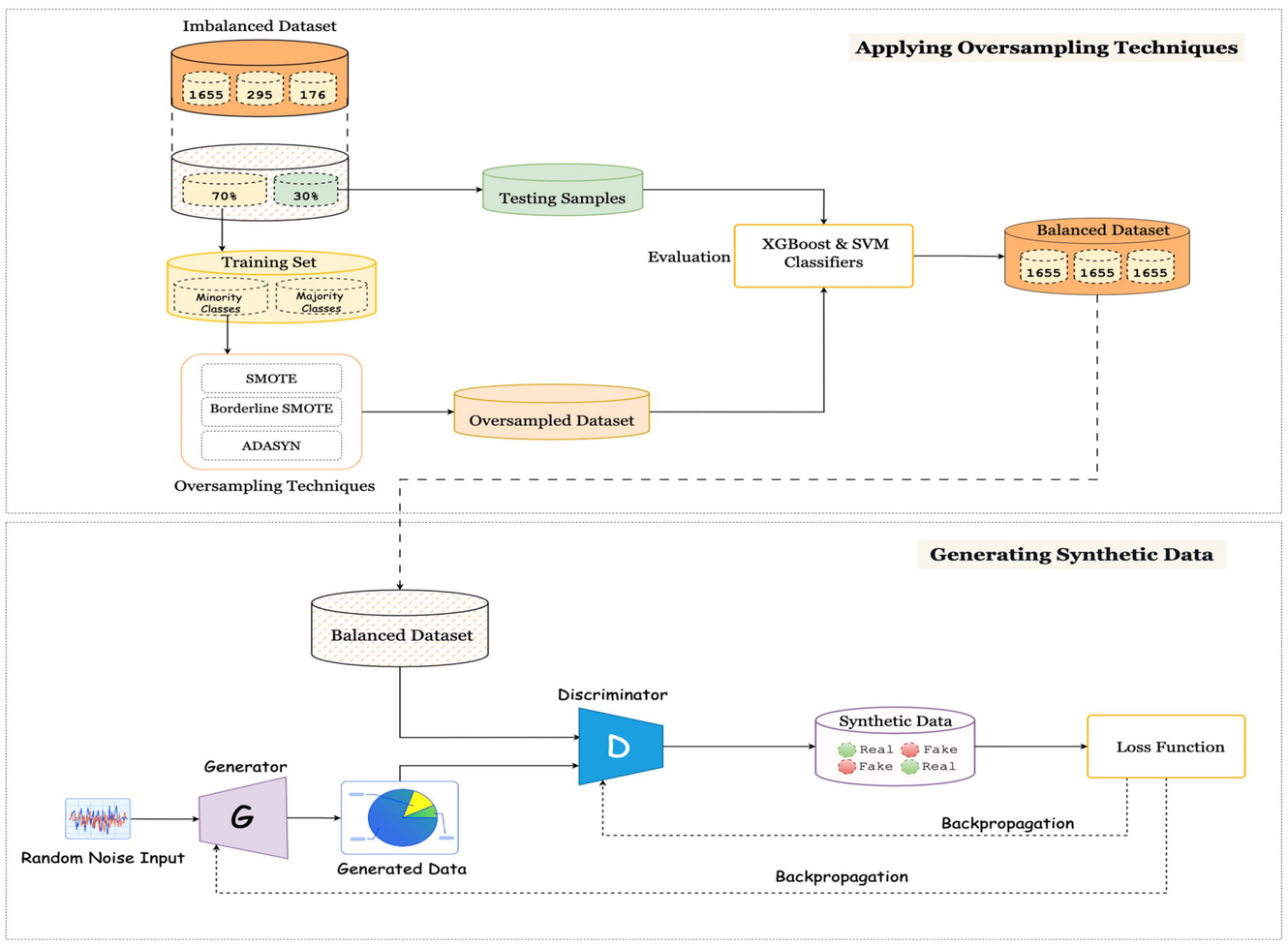
Addressing Class Imbalance in Medical Datasets using Oversampling Techniques and Generative Adversarial Networks

Halal Abdulrahman Ahmed[1], Juan A. Nepomuceno[1], Belén Vega-Márquez[1], and Isabel A. Nepomuceno-Chamorro[1]
[1]Dpto de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Sevilla, Spain.
halahmabd@alum.us.es

Abstract

This study employs three oversampling techniques, SMOTE, ADASYN, and BorderlineSMOTE, to solve the class imbalance in the Fetal Cardiotocography (CTG) dataset. After balancing minority classes in the (CTG) dataset, we employ two variations of Generative Adversarial Networks (GANs), including CGAN and CTGAN, to generate tabular synthetic data. Moreover, for the purpose of performance evaluation, we use XGBoost and SVM classifiers. This research sheds light on the potential of oversampling techniques and GANs to mitigate class imbalances and solve the lack of enough data.

Methods



Highlights

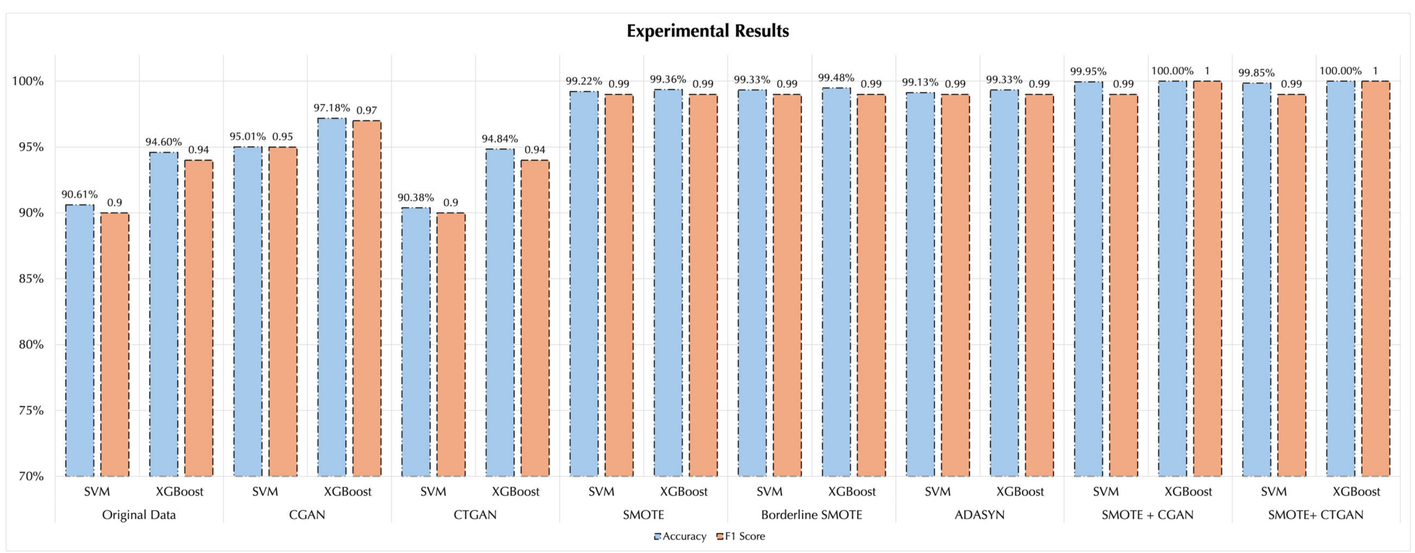
Oversampling techniques and GAN models have proven to be effective in addressing class imbalances, often leading to improved classifier performance. However, their success can depend on the specific dataset and implementation.

In order to balance the dataset, our future work will create an oversampling technique that combines traditional oversampling techniques with generative deep learning models.

Our previous study (Ahmed et al., 2024) explored synthetic data generation using various GAN models, yielding promising results. The current study integrates GAN models with traditional oversampling techniques. Initially, we applied oversampling methods such as SMOTE, Borderline SMOTE, and ADASYN to balance the dataset. After that, we used the balanced dataset as input for GAN models (CGAN, CTGAN) to generate synthetic data.

Experimental Results

We evaluated the performance of two classifiers, SVM and XGBoost, on original, oversampled, and balanced datasets. The experimental results showed that data balancing methods significantly enhance classifier performance. The combination of SMOTE and CGAN or CTGAN achieved perfect scores with XGBoost.



References

Ahmed, H. A., Nepomuceno, J. A., Vega-Márquez, B., & Nepomuceno-Chamorro, I. A. (2024). Enhancing healthcare data privacy with generative adversarial networks: A comparative study on synthetic data generation for medical datasets. Manuscript under revision.

UNIVERSIDAD DE SEVILLA
JABI 2024 GRANADA, 24, 25 Y 26 DE JUNIO
Junta de Andalucía