Data Synthesis on Unbalanced Datasets Using Machine Learning

Puumaaya Tahiru, Isabella Sunderman and James Crabb

INTRODUCTION

- In recent years, machine learning has emerged as a powerful tool in cybersecurity research, offering innovative solutions to address the dynamic and sophisticated nature of cyber threats. With the ever-evolving landscape of cyber threats, traditional rule-based approaches to cybersecurity are proving inadequate. By leveraging advanced algorithms and large datasets, machine learning models can detect and prevent cyber-attacks in real-time, thus enhancing the overall security posture of organizations.
- Data imbalance poses a significant challenge in cybersecurity research. In many cases, cybersecurity datasets are highly imbalanced, with a few instances of malicious activities overshadowed by many benign instances. This imbalance can lead to biased models that perform poorly in detecting real threats. To address this issue, researchers and cybersecurity professionals often resort to synthetic data generation techniques.
- In our research, we explore how balanced datasets ensure that machine learning models can accurately identify potential security risks while minimizing false positives. We also discuss the different methods we used to generate synthetic data from existing cybersecurity datasets. By creating balanced datasets through synthetic data generation, we aim to improve the performance and reliability of machine learning models in cybersecurity applications, ultimately enhancing the security posture of organizations in the face of evolving cyber threats.

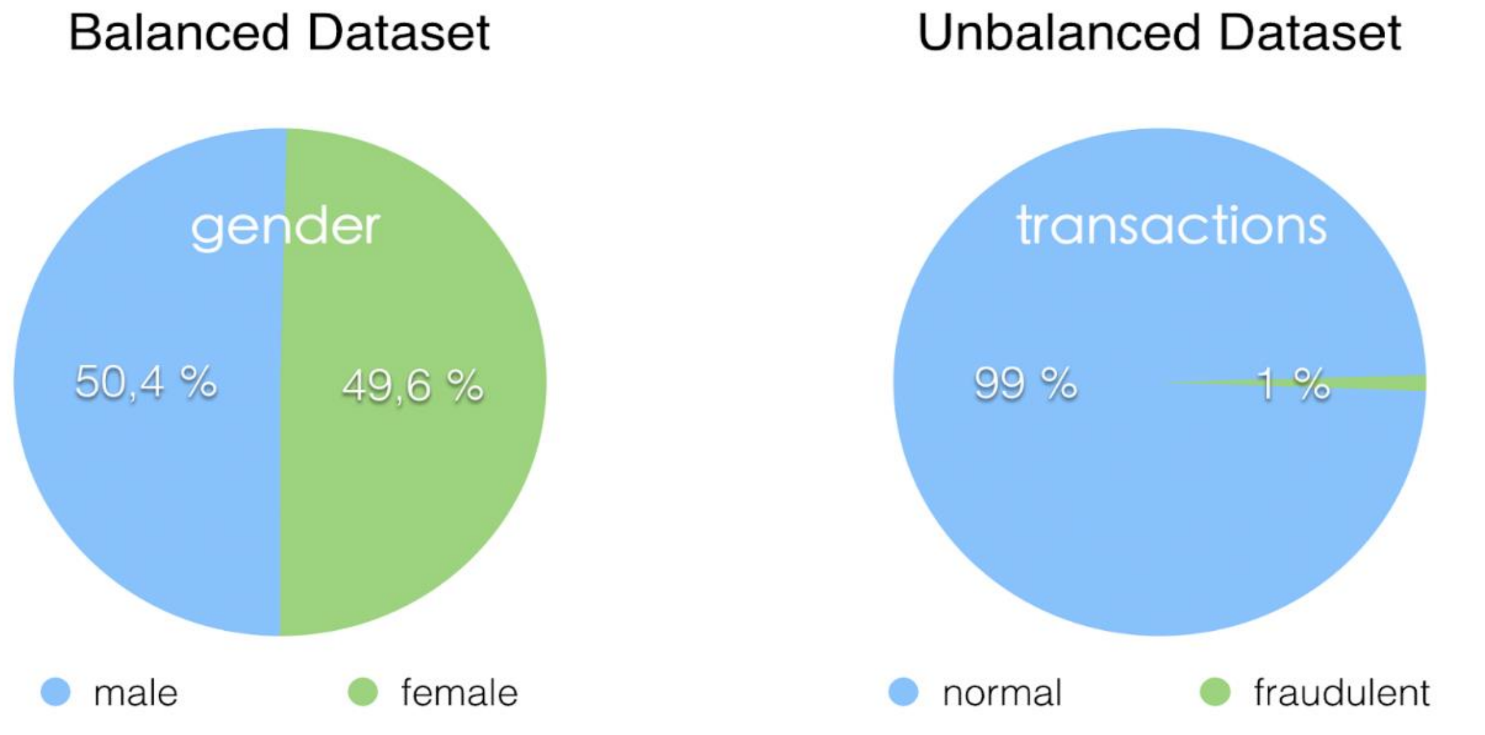

Figure 1: Graphic of the difference between a balanced and unbalanced dataset. Source: unbalanced-datasets-what-to-do-144e0552d9cd

METHODS

- SMOTE- Synthetic Minority Over-sampling Technique creates synthetic samples from a dataset utilizing the minority samples features and attributes. The SMOTE algorithm selects a K^{th} nearest neighbor to a point in a data set to generate a new sample between the two data points[1].
- GAN- Generative Adversarial Networks uses a generative (G) and discriminative (D) network that work against each other. The generative network synthesizes data that is close to the given dataset to lower the confidence of the discriminatory network. The discriminatory network works to guess whether the data is from the original data set or not with high confidence[2].
- Random Forest and K^{-}Nearest Neighbors Classifiers were both fitted with data from a training set of samples and the synthetic samples produced using SMOTE and GAN separately.
- Random Forest Classifier is a commonly used machine learning technique that uses a series of decision trees and averaging to improve accuracy of predictions[3].
- K^{-}Nearest Neighbors Classifier learns from a given data set by using the *k* nearest neighbors of a data point and an instance-based approach to the data[4]

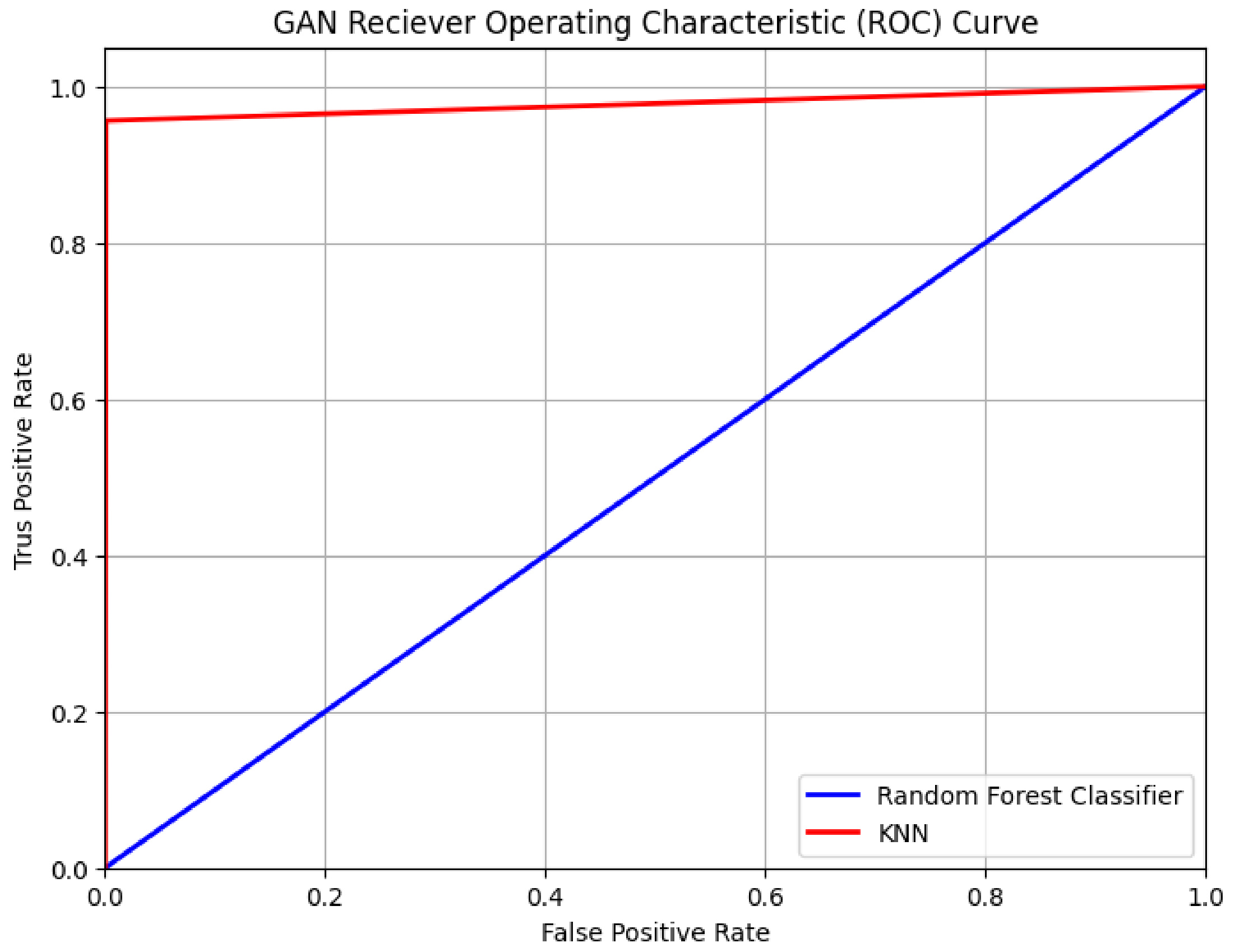

Figure 2: ROC curve of GAN data using Random forest classifier and K-nearest neighbors' classifier

DATA SYNTHESIS

- For our research, we primarily focused on utilizing a credit card fraud dataset. Initially, we downloaded the dataset and performed data preprocessing, ensuring that there were no unnecessary columns, missing values, or empty entries. Following this, we split the data into training and test sets, using a 75% to 25% ratio, respectively.
- To address the issue of class imbalance, we applied two different data augmentation techniques: SMOTE (Synthetic Minority Over-sampling Technique) and GAN (Generative Adversarial Network). These techniques allowed us to generate synthetic data from the credit card dataset, thereby balancing the distribution of fraudulent and non-fraudulent transactions.
- After generating synthetic data using both SMOTE and GAN algorithms, we sorted the new datasets and compared the performance of each algorithm. Specifically, we evaluated the ROC scores of classifiers trained on the original and synthetic datasets to assess the effectiveness of each data augmentation technique in improving the model's ability to detect fraudulent transactions.

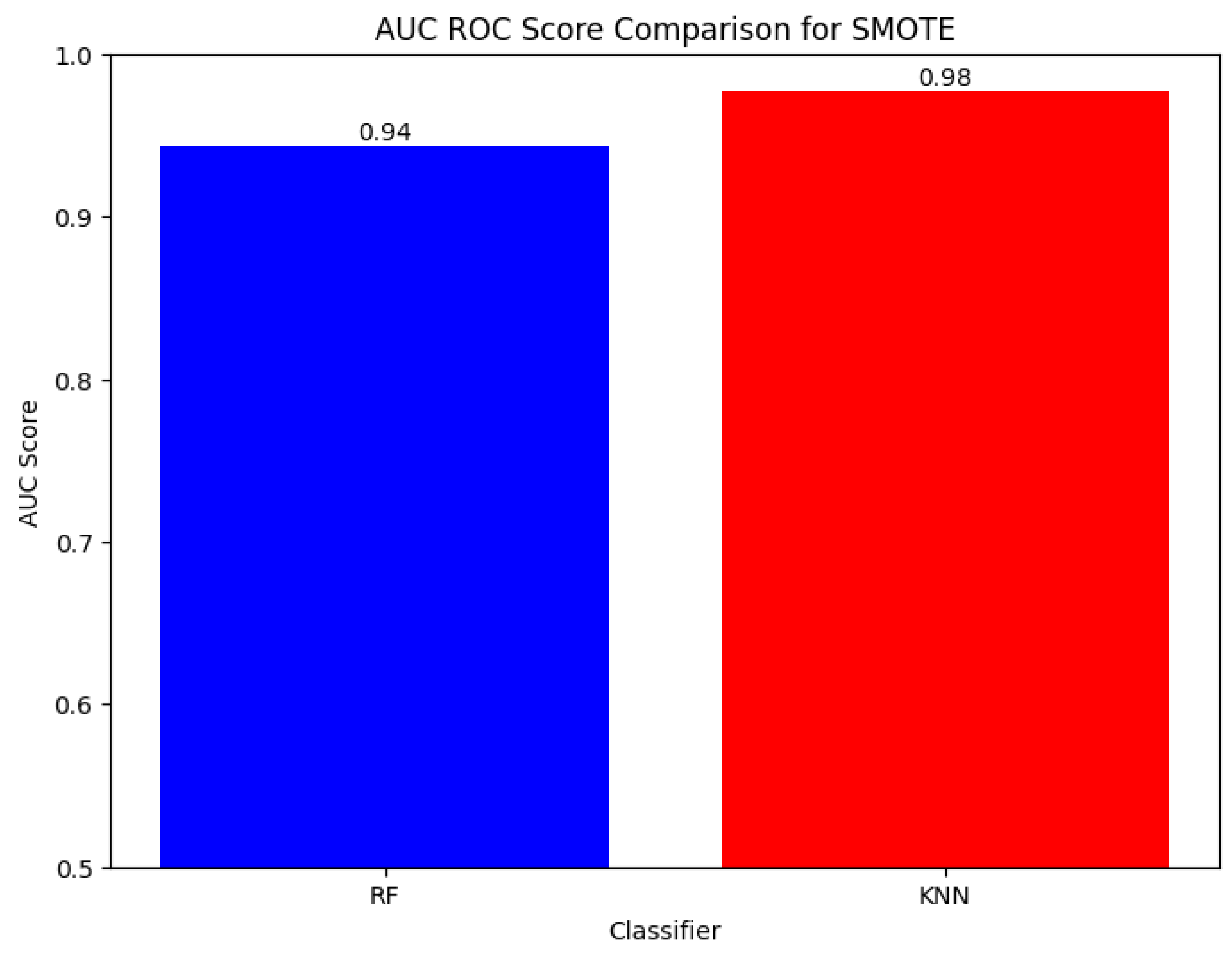

Figure 3: Performance of RF and KNN on SMOTE data

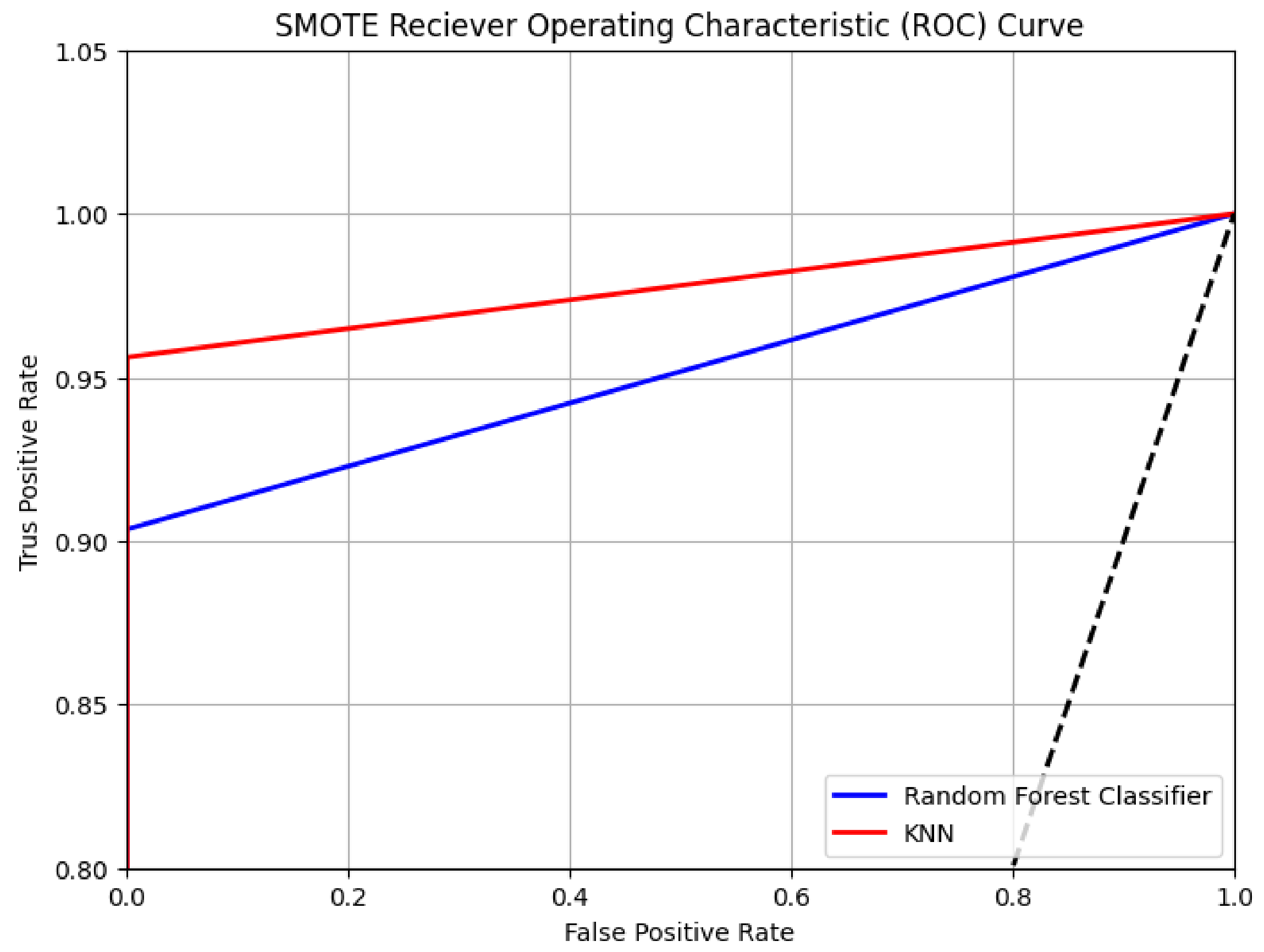

Figure 4: ROC curve of SMOTE data using Random forest classifier and K-nearest neighbors classifier

REFERENCES

1. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002), 321–357.
2. Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. NeurIPS, 2019 (accessed May 13, 2024)
3. "1.11. ensembles: Gradient boosting, random forests, bagging, voting, stacking," scikit, https://scikit-learn.org/stable/modules/ensemble.html#forest (accessed May 13, 2024).
4. J. Vanderplas, "1.6. nearest neighbors," scikit, https://scikit-learn.org/stable/modules/neighbors.html#classification (accessed May 13, 2024).

ACKNOWLEGEMENTS

This work is supported by funding for the VICEROY Northwest Institute for Cybersecurity Education and Research (CySER) provided by The Office of the Undersecretary of Defense for Research and Engineering, in collaboration with the Air Force Research Laboratory and Griffiss Institute.

WASHINGTON STATE UNIVERSITY