A Comparative Study of Traditional Methods like ROS, RUS, SMOTE and their hybrid combinations with GAN Based Methods like CGAN in Dealing with Classical Imbalance Problem.

1st Kaushik Raj Ghimire
Dept. of CSE(AIML)
Ramaiah Institute of Technology
Bangalore, India
ghimirekaushiraj116@gmail.com

2nd Bishal Babu Rajbanshi
Dept. of CSE(AIML)
Ramaiah Institute of Technology
Bangalore, India
bishalraj10.np@gmail.com

3rd Rajan Lal Karna
Dept. of CSE(AIML)
Ramaiah Institute of Technology
Bangalore, India
rajanlalkarna.np@gmail.com

***Abstract*—The classical imbalance problem poses significant challenges for machine learning models, leading to poor performance on minority class predictions in real-world applications such as fraud detection and medical diagnosis. This study provides a comparative analysis of traditional resampling methods, including Random Oversampling (ROS), Random Undersampling (RUS), and Synthetic Minority Oversampling Technique (SMOTE), alongside hybrid techniques such as SMOTE+ENN, SMOTE+Tomek Links, and SMOTE-NC+RUS. Furthermore, advanced Generative Adversarial Networks (GANs), specifically Conditional GAN (CGAN), were employed to generate diverse, high-quality synthetic samples. Experimental results, evaluated using metrics like accuracy,precision,recall,F1-score and ROC-AUC score, revealed the limitations of traditional methods like overfitting in ROS and information loss in RUS, while hybrid methods achieved better class balance and performance. Among all, CGAN outperformed traditional approaches, demonstrating its potential to address class imbalance effectively. This study bridges classical and deep learning techniques, offering insights into optimal methods for imbalanced datasets.**

***Index Terms*—Classical Imbalance, Resampling Techniques, CGAN, XGBoost, SMOTE**

I. Introduction

The class imbalance problem is a pervasive challenge in machine learning, where models often struggle to accurately predict the minority class due to skewed data distributions [1]. This issue is particularly critical in domains like fraud detection, medical diagnosis, and anomaly detection, where accurate minority class predictions are vital. Conventional resampling techniques, such as Random Oversampling (ROS), Random Undersampling (RUS), and the Synthetic Minority Oversampling Technique (SMOTE), have been extensively employed to address this issue by rebalancing datasets [2]. However, these methods face inherent limitations, such as overfitting in ROS and information loss in RUS, which hinder their effectiveness [3].

Hybrid approaches, including SMOTE combined with Edited Nearest Neighbors (SMOTE+ENN), Tomek Links (SMOTE+Tomek), and SMOTE-Nominal Continuous with Random Undersampling (SMOTE-NC+RUS), have emerged as enhanced alternatives, addressing noise and class boundary issues. These methods strive to balance precision and recall, achieving better performance on imbalanced datasets [4].

Recently, advanced Generative Adversarial Networks (GANs) have gained traction as a transformative solution to class imbalance. Conditional GANs (CGANs), in particular, generate high-quality, diverse synthetic samples tailored to the minority class, surpassing traditional methods in capturing underlying data distributions [5] [6].

This paper provides a comprehensive comparative analysis of these traditional, hybrid, and GAN-based methods. Through rigorous experimentation and evaluation using metrics like accuracy, precision, recall, F1-score and ROC-AUC, this study highlights the strengths and limitations of each approach. The findings underscore the potential of GAN-based resampling as a cutting-edge solution for tackling imbalanced learning problems in real-world applications.

II. Related Work

The paper [7] addressed class imbalance problem by integrating SMOTE with Convolutional Neural Networks (CNNs), applying their methods to 24 datasets, including KEEL, breast cancer, and Z-Alizadeh Sani. They study used both oversampling techniques like ROS and SMOTE and undersampling methods like RUS and Tomek Links, integrating them with deep learning models. To improve learning, they employed focal loss and validated their approach using multiple metrics, including AUC-ROC, F1-score, and Kappa. The results showed strong performance while addressing challenges like overfitting and information loss, offering a reliable framework for handling imbalanced classification problems.

In the study conducted by [8], compared three resampling methods—Random Oversampling (ROS), Random Undersampling (RUS), and a hybrid SMOTE-NC+RUS approach on High School Longitudinal Study of 2009 dataset. They highlighted RUS's information loss and ROS's overfitting in extreme scenarios, emphasizing the importance of selecting resampling techniques based on imbalance ratios to improve model performance.

The literature [9] proposed advanced oversampling techniques, Distance-based SMOTE (D-SMOTE) and Bi-phasic SMOTE (BP-SMOTE), integrated with stacking ensemble frameworks like Stacked CNN and Stacked RNN to tackle class imbalance issues in healthcare. These hybrid approaches achieved significantly higher accuracies of 96–97%, surpassing traditional methods. The models demonstrated good performance by mitigating challenges such as class overlap and information loss.

The authors of [10] evaluated six machine learning models— Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM), and XGBoost on imbalanced and SMOTE-balanced datasets for predicting hospital mortality in trauma patients. SMOTE techniques like Borderline-SMOTE, SMOTE-NC, and SVM-SMOTE significantly improved sensitivity, AUC, and overall performance. ANN and RF performed best with SMOTE, while XGBoost excelled with SMOTE-NC, highlighting the effectiveness of SMOTE in addressing class imbalance.

The study [11], [12], [13] addressed class imbalance with traditional methods like SMOTE and advanced GAN-based techniques. While SMOTE balanced classes but struggled with data overlap, GAN produced high-quality, diverse samples that preserved minority class distribution. GAN-based methods, especially CTGAN, consistently outperformed traditional approaches by reducing overfitting, handling high-dimensional data effectively, and improving accuracy, precision, and F1-scores, highlighting GAN's advantage in managing imbalanced data.

III. Methodology

A. Dataset Preparation

The dataset used in this study is the Credit Card Fraud Detection dataset from Kaggle, containing 5,000 records and 17 features with a significant class imbalance (91% non-fraudulent and 9% fraudulent cases), resulting in an imbalance ratio of around 9:1. Preprocessing steps included handling missing values, removing outliers, and standardizing the independent features using StandardScaler to ensure the data's readiness for model training and evaluation [7]. The target variable was encoded into binary classes, preserving the dataset's integrity for effective analysis.

B. Exploratory Data Analysis (EDA) and Feature Selection

EDA and feature selection were performed to gain insights into the dataset's structure and optimize feature usage [14]. Features with correlations greater than 0.9 were removed using a correlation matrix to address multicollinearity, reduce redundancy, and improve model efficiency, particularly for tree-based models. Visualization techniques, including correlation heatmaps, class distribution histograms, and pair plots, were employed to analyze feature relationships and highlight the class imbalance. Stratifying the target variable during sampling ensured balanced representation across training and testing sets. [15].

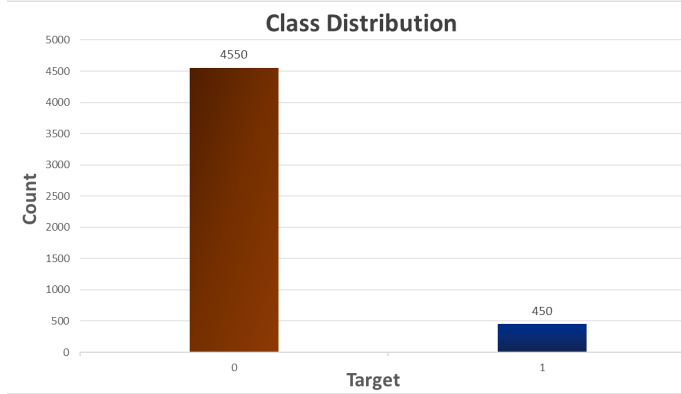

Fig. 1. Class Distribution

C. Train-Test split

The dataset was divided into training and testing subsets using an 80:20 split. This split ensured sufficient data for model training while reserving a separate portion for unbiased performance evaluation.

D. Baseline Model Implementation

To establish baseline performance metrics, four supervised machine learning models—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and XGBoost—were implemented. A robust cross-validation approach was employed to mitigate overfitting and ensure reproducibility. The models achieved accuracy,precision,recall, F1-score and ROC-AUC scores as Logistic Regression:(0.956, 0.901, 0.541, 0.676, 0.701), Decision Tree: (0.943, 0.741, 0.505, 0.601, 0.779), Random Forest:(0.958, 0.977, 0.517, 0.676, 0.903), and XGBoost:(0.959, 0.879, 0.6, 0.713, 0.933) respectively, highlighting XGBoost as the best-performing model across overall performance metrics. The F1-score was chosen as the primary evaluation metric due to its ability to balance precision and recall, effectively balancing the trade-off between false positives and false negatives, providing a more comprehensive measure of performance for imbalanced datasets where misclassification of the minority class can have a greater impact. XGBoost was subsequently selected as the candidate model for resampling experiments based on its superior baseline results [16].

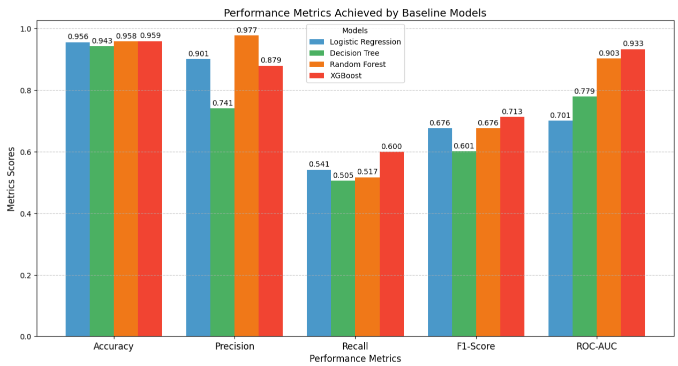

Fig. 2. Performance Metrics Achieved by Baseline Models

E. Traditional Resampling Methods

1) Random Over-Sampling (ROS): ROS duplicates existing minority class samples randomly to balance the dataset. While simple and effective, it may lead to overfitting because identical samples are added without introducing any variation.

2) Random Under-Sampling (RUS): RUS reduces the majority class by randomly removing samples, aiming to balance the classes. Although it prevents overfitting, this approach risks losing potentially valuable information from the majority class.

3) Synthetic Minority Oversampling Technique (SMOTE): SMOTE generates synthetic samples for the minority class using the k-nearest neighbors algorithm. It interpolates between minority samples, creating new data points that resemble the existing distribution, which reduces the risk of overfitting compared to ROS.

4) Hybrid Methods: Hybrid methods combine both oversampling and undersampling techniques to enhance the strengths of each while mitigating their weaknesses. These methods aim to create a more balanced dataset without introducing significant information loss or overfitting.

a) SMOTE-ENN (Edited Nearest Neighbors): This hybrid technique combines SMOTE with ENN. After synthetic samples are generated by SMOTE, ENN removes samples that are ambiguous or incorrectly classified by their nearest neighbors, ensuring cleaner and more accurate boundaries between classes.

b) SMOTE-Tomek Links: After applying SMOTE, Tomek Links remove overlapping or noisy samples between the majority and minority classes. This ensures that the dataset is not only balanced but also free from borderline conflicts.

c) SMOTE-NC + RUS: SMOTE-NC is an extension of SMOTE for mixed datasets containing both numerical and categorical features. It oversamples the minority class while considering categorical feature relationships. Combining it with RUS removes excess majority samples, striking a balance between diversity and simplicity [4].

Hybrid resampling techniques, such as SMOTE + ENN, SMOTE + Tomek Links, and SMOTE-NC + RUS, consistently delivered superior results by addressing key challenges in imbalanced datasets. These methods effectively tackled issues of noise and class overlap by removing ambiguous and conflicting samples, leading to cleaner decision boundaries and better model generalization. By combining the strengths of oversampling, which enhances minority class representation, and under-sampling, which reduces majority class dominance, these techniques achieved an optimal balance without introducing excessive complexity. SMOTE-NC further enriched the process by synthesizing realistic samples for both categorical and numerical features, ensuring the integrity of the dataset structure. This synergy minimized redundancy, mitigated noise, and preserved critical information, overcoming the limitations of traditional methods like ROS, RUS, and SMOTE. Consequently, XGBoost trained on hybrid-resampled data achieved significantly higher F1-scores, demonstrating the robustness and effectiveness of these methods in handling class imbalance.

XGBoost, renowned for its scalability and robustness, was fine-tuned using Randomized Grid Search to optimize performance on the imbalanced dataset. Key hyperparameters included learning_rate (0.01–0.3) for gradual learning, min_child_weight (1–4) to prevent overfitting, gamma (0–10) to control model complexity, and subsample (0.7–0.9) and colsample_bytree (0.7–0.9) to enhance generalization by introducing randomness. The max_depth (3–9) parameter balanced complexity and overfitting. Randomized Grid Search efficiently explored the hyperparameter space while conserving computational resources, with F1-score as the evaluation metric to ensure an optimal balance of precision and recall.

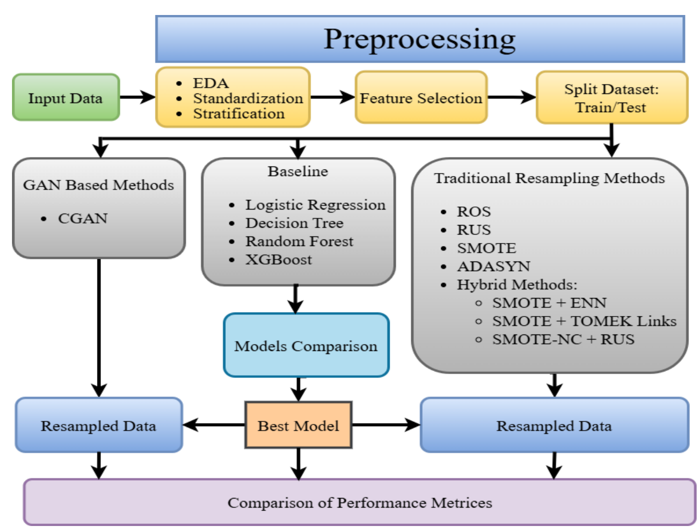

Fig. 3. Flow diagram of the project

F. GAN-based Methods

1) Conditional Generative Adversarial Network (CGAN): The Conditional Generative Adversarial Network (CGAN) was designed with two neural networks, the Generator and the Discriminator, each playing a critical role in the model's performance. The Generator took a random noise vector of size 100 (latent dimension) and concatenated it with the class label to form the input. This input was processed through three fully connected layers with ReLU activation functions, progressively generating synthetic samples. A Sigmoid activation function in the final layer ensured that the output data conformed to the normalized range of the original features. The Discriminator, tasked with evaluating the authenticity of the samples, combined the input data with their respective class labels. These concatenated inputs passed through several fully connected layers with LeakyReLU activation functions, offering improved gradient flow, and a Sigmoid activation function at the end classified the inputs as real or synthetic.

The CGAN was trained using Binary Cross-Entropy Loss for both networks in an alternating optimization procedure. The Discriminator minimized the loss for distinguishing real samples, labeled as 111, from synthetic samples, labeled as 000. Simultaneously, the Generator optimized its parameters to produce synthetic samples capable of "fooling" the Discriminator into classifying them as real. Both networks were trained with the Adam optimizer, using a learning rate of 0.0001,0.0001,0.0001 and hyperparameters β_1=0.5 and β_2=0.999. To incorporate class information, the CGAN used a label conditioning approach by concatenating noise vectors with class labels, with slight noise added to the labels during training to enhance robustness and generalization.

After 200 epochs of training, synthetic data was generated for the minority class by sampling random noise vectors from a standard normal distribution, concatenating these vectors with the minority class label, and feeding the inputs into the Generator. The generated samples were inverse-transformed using the original feature scaler to restore them to the original data range. These synthetic samples were then combined with the original dataset, creating a balanced dataset to address the class imbalance issue. The enriched dataset was used to train an XGBoost classifier with hyperparameter tuning to maximize the F1-score, demonstrating significant improvements in minority class performance and highlighting the effectiveness of CGAN-generated synthetic data in addressing class imbalance [17] [18].

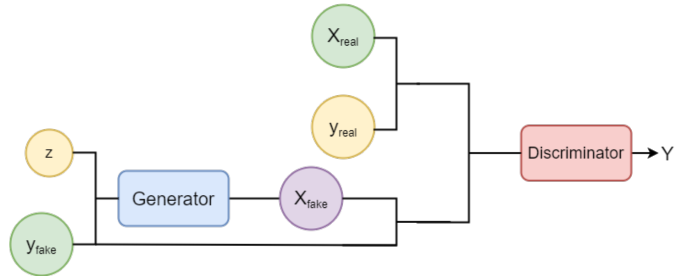

Fig. 4. Conditional GAN

IV. Results and Discussions

A. Traditional Resampling Techniques

1) Random Oversampling (ROS): ROS improved recall to 0.6941, resulting in a higher F1-score of 0.7763, with precision at 0.8806. The method achieved an accuracy of 0.966 and a ROC-AUC of 0.9632. However, it suffered from overfitting, as duplicating minority class samples without introducing new diversity caused the model to focus too heavily on the minority class.

2) Random Undersampling (RUS): RUS boosted recall to 0.8471, but the significant reduction in majority class samples caused information loss, leading to low precision (0.4311) and an F1-score of 0.5714. The method achieved an accuracy of 0.892 and a ROC-AUC of 0.9315, but its limitations were evident in effectively handling class imbalance.

3) SMOTE: SMOTE synthesized new samples for the minority class, achieving an F1-score of 0.7895, with precision at 0.8955 and recall at 0.7059. The method recorded an accuracy of 0.968 and a ROC-AUC of 0.9658, showing balanced improvements across all metrics.

4) ADASYN: ADASYN improved class balance by adaptively generating synthetic samples for harder-to-learn areas. It achieved an accuracy of 0.964, F1-score of 0.785, precision of 0.795, recall of 0.776, and a ROC-AUC of 0.967, demonstrating its ability to focus on challenging decision boundaries while avoiding overfitting.

5) Hybrid Approaches:

a) SMOTE+ENN: This method combined SMOTE with Edited Nearest Neighbors (ENN) to remove noisy samples. It achieved an F1-score of 0.8114, precision of 0.7889, recall of 0.8353, accuracy of 0.967, and a ROC-AUC of 0.9678, indicating its ability to handle noisy data while maintaining balanced class performance.

b) SMOTE+TOMEK: By integrating SMOTE with Tomek Links, this method removed borderline majority samples. It achieved an F1-score of 0.8075, precision of 0.8553, recall of 0.7647, accuracy of 0.969, and a ROC-AUC of 0.9663, highlighting its strength in refining class boundaries.

c) SMOTE-NC+RUS: This hybrid method combined SMOTE-NC for handling categorical data with RUS for balancing majority samples. It was the best traditional method, achieving an F1-score of 0.82, precision of 0.8214, recall of 0.8118, accuracy of 0.969, and a ROC-AUC of 0.9713. It effectively balanced precision and recall while minimizing overfitting and information loss.

B. GAN-Based Methods

1) CGAN: CGAN significantly outperformed traditional methods by generating high-quality, diverse synthetic samples for the minority class. It achieved an accuracy of 0.98, F1-score of 0.9015, precision of 0.9520, recall of 0.8561, and a ROC-AUC of 0.9731. This method successfully captured the underlying distribution of the minority class, allowing for superior generalization and robust performance.

TABLE I
RESULTS FOR EACH RESAMPLING TECHNIQUE

Resampling Techniques	**Performance Metrics**				
	Accuracy	***Precision***	***Recall***	***F1 Score***	***ROC-AUC Score***
Baseline	0.959	0.879	0.600	0.713	0.933
ROS	0.966	0.880	0.694	0.776	0.963
RUS	0.892	0.431	0.847	0.571	0.931
SMOTE	0.968	0.895	0.705	0.789	0.965
ADASYN	0.964	0.795	0.776	0.785	0.967
SMOTE+ENN	0.967	0.788	0.835	0.811	0.967
SMOTE+TOMEK	0.969	0.855	0.764	0.807	0.966
SMOTE-NC+RUS	0.969	0.821	0.811	0.820	0.971
CGAN	0.980	0.987	0.885	0.933	0.973

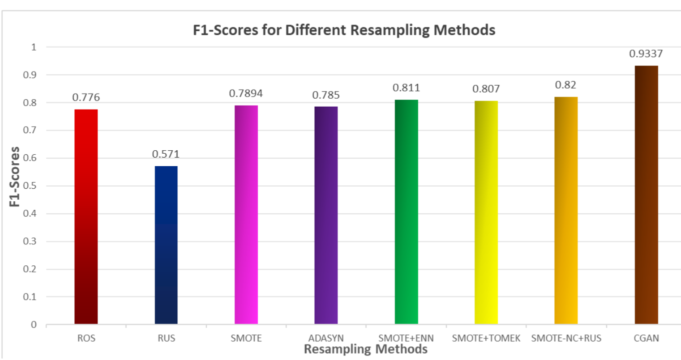

Fig. 5. F1-Scores for Different Resampling Methods

V. CONCLUSION

Therefore, we can conclude that while ROS suffered from overfitting and RUS from information loss, hybrid methods like SMOTE-NC+RUS and SMOTE+ENN demonstrated a balanced approach by effectively improving precision and recall. Among these, SMOTE-NC+RUS emerged as the most reliable traditional method. However, CGAN surpassed all traditional techniques by generating diverse, high-quality synthetic samples that significantly mitigated class imbalance. Its superior F1-score, ROC-AUC score, accuracy, and precision-recall balance highlight the potential of GAN-based methods as a transformative solution for addressing real-world imbalanced learning challenges.

REFERENCES

[1] Liang, Zefeng, et al. "Adaptive fusion based method for imbalanced data classification." Frontiers in Neurorobotics 16 (2022): 827913.

[2] Lin, Cian, Chih-Fong Tsai, and Wei-Chao Lin. "Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study." Artificial Intelligence Review 56.2 (2023): 845-863.

[3] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[4] Wongvorachan, Tarid, Surina He, and Okan Bulut. "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining." Information 14.1 (2023): 54.

[5] Dam, Tanmoy. Developing generative adversarial networks for classification and clustering: Overcoming class imbalance and catastrophic forgetting. Diss. UNSW Sydney, 2022.

[6] Fajardo, Val Andrei, et al. "On oversampling imbalanced data with deep conditional generative models." Expert Systems with Applications 169 (2021): 114463.

[7] Joloudari, Javad Hassannataj, et al. "Effective class-imbalance learning based on SMOTE and convolutional neural networks." Applied Sciences 13.6 (2023): 4006.

[8] Wongvorachan, Tarid, Surina He, and Okan Bulut. "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining." Information 14.1 (2023): 54.

[9] Sowjanya, A. Mary, and Owk Mrudula. "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms." Applied Nanoscience 13.3 (2023): 1829-1840.

[10] Hassanzadeh, Roghayyeh, Maryam Farhadian, and Hassan Rafieemehr. "Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms." BMC medical research methodology 23.1 (2023): 101.

[11] Park, Jungryeol, Sundong Kwon, and Seon-Phil Jeong. "A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks." Journal of Big Data 10.1 (2023): 36.

[12] Eom, Gayeong, and Haewon Byeon. "Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique." Mathematics 11.16 (2023): 3605.

[13] Ayoub, Shahnawaz, et al. "Adversarial approaches to tackle imbalanced data in machine learning." Sustainability 15.9 (2023): 7097.

[14] Gupta, Palak, et al. "Unbalanced credit card fraud detection data: a machine learning-oriented comparative study of balancing techniques." Procedia Computer Science 218 (2023): 2575-2584.

[15] Widodo, Akdeas Oktanae, Bambang Setiawan, and Rarasmaya Indraswari. "Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE." Procedia Computer Science 234 (2024): 578-583.

[16] Riyanto, Slamet, et al. "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification." International Journal of Advanced Computer Science

and Applications 14.6 (2023).

[17] Sampath, Vignesh, et al. ”A survey on generative adversarial networks for imbalance problems in computer vision tasks.” Journal of big Data 8 (2021): 1-59.

[18] Engelmann, Justin, and Stefan Lessmann. ”Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning.” Expert Systems with Applications 174 (2021): 114582.