Bulletin of Business and Economics, 13(3), 385-395
https://bbejournal.com
https://doi.org/10.61506/01.00505

Enhanced Wasserstein Generative Adversarial Network (EWGAN) to Oversample Imbalanced Datasets

Muhammad Hassan Ajmal Hashmi[1], Muhammad Ashraf[2], Saleem Zubair Ahmad[3], Muhammad Waseem Iqbal[4], Adeel Hamid[5], Dr. Abid Ali Hashmi[6], Muhammad Ameer Hamza[7]

Abstract

This paper examines WGAN as a more advanced technique for addressing imbalanced data sets in the context of machine learning. A variety of domains, including medical diagnosis and image generation, are affected by the problem of imbalanced datasets since it is essential to represent the minority class to train a satisfactory model and create various types of data. To overcome these challenges WGAN uses some features such as; Residual connections in the critic network, better sampling for minority classes, and some noise and sample reshaping. These innovations contribute to the increased stability of the model, the quality of synthetic data, and the distribution of classes in a dataset. The comparative analysis of WGAN with basic GAN and Improved GAN has shown the effectiveness of the given algorithm in terms of producing high-quality diversified synthetic data that is closer to the real data distribution. The study identifies the future research direction of WGAN in enhancing machine learning based on reliable and diverse synthesized data, providing new insights and directions for future studies and practical applications in tackling data imbalance issues.

Keywords: WGAN, Imbalanced Data, Synthetic Data, Machine Learning, Cancer Diagnosis, Data Sampling, Model Stability, Data Generation, GAN Models

1. Introduction

1.1. Background

Handling imbalance data is a major problem in machine learning and has a huge impact on the predictive models in many areas. In other words, in imbalanced datasets, there are more instances of a specific class (Suh et al., 2021). In comparison, the rest of them have much fewer instances and this contributes to models that have poor capability, especially when it comes to making predictions on the minority class. This issue is particularly alarming in diagnostics, specifically in assigning rare but severe conditions like malignant tumors when reaching an accurate diagnosis is vital to the proper treatment and care of the patient.

In certain applications like image generation and diagnosis, it's crucial to ensure that all classes are given an equal proportion. Current gradual approaches trained on imbalanced data can hardly predict the minority classes which leads to a low performance level of the model. This requires superior methods to improve the process of displaying and combining data for training and generalization of the models in the other classes (Zhang, Liu, Luan, & Sun, 2017).

In medical diagnosis, datasets usually consist of a greater number of instances that are labeled as non-cancerous, thus affecting the distribution. In the case of such skewed datasets, the old generated machine learning algorithms that fit on such data tend to perform poorly when it comes to labeling the minority classes right. This is also true for other fields like image generation where it is crucial to have the data diverse and balanced to generate synthetic images which are as real-looking as possible about real-life situations (Ali, Dastgir, Iqbal, Anwar, & Faheem, 2023).

The problem of data distribution imbalance is not a newly emerged challenge in the field of machine learning and data analysis but has been observed recently in many applications. When there are subsets of data with considerably fewer data points than the others, training classifiers to recognize the patterns can be quite difficult. Most of the previous algorithms for machine learning are inclined towards the majority class, which in most cases proves detrimental to the minority class. To overcome this problem, different oversampling methods have been proposed to create new synthetic samples of the minority class and consequently increase its proportion in the dataset. As for the most perspective methods in the given field, it is possible to note the Generative Adversarial Networks (GANs) (Goodfellow et al., 2020).

Generative Adversarial Networks is a model that contains two neural networks popularly known as the Generator and the Discriminator, and they are trained at the same time in an antagonistic manner. The goal of the generative model is to examine the collection set of training examples and learn the probability distribution that properly generated them. GAN is then fully able to generate more examples from the proposed and estimated probability distribution. The Generator's purpose is to create realistic synthetic data; on the other hand, the Discriminator is aimed at predicting whether the data is real or fake (Arjovsky, Chintala, & Bottou, 2017).

The next advancement over the basic GANs is the Improved Generative Adversarial Network (IGAN), where improvements are made to the quality and the variety of the data being generated. To mitigate the effects of training instability and derived synthetic data low fidelity, IGAN provides architectural and training improvements. Although IGAN can overcome some of the obstacles that stem from standard GANs, it might also struggle with some more issues typical for class imbalance and data distribution that WGAN claims to be less sensitive.

To solve these problems, the Wasserstein Generative Adversarial Network (WGAN) was introduced. WGAN also proposes the usage of a new distance, the Earth Mover's Distance (EMD), or the Wasserstein distance between the real and generated data distributions (Lee, Li, & Li, 2023). This modification also helps in improving the stability of the training process and it also increases the convergence properties. The Critic in WGAN replaces the Discriminator in I-GAN and is intended to be trained to estimate the Wasserstein distance between the target and generated data distributions.

[1] CEO Tech Solutions Lahore-54000 Pakistan, Faculty of Computer Science, KIPS College Lahore-5400 Pakistan, hassan.ajmal123@gmail.com
[2] IT Department ,Gulab Devi Teaching Hospital Lahore 54000, Pakistan, itdept@gulabdevi.org
[3] Department of Software Engineering, Superior University Lahore-54000, Pakistan, saleem.zubair@superior.edu.pk
[4] Department of Software Engineering, Superior University Lahore-54000, Pakistan, waseem.iqbal@suprior.edu.pk
[5] Faculty of Computer Science, Virtual University Lahore-5400, Pakistan, adeel.hamid50@gmail.com
[6] Educational Complex Lahore-54000 Pakistan, project.dir.aamc@gmail.com
[7] Department of Computer Science, Superior University Lahore-54000, Pakistan, alihamza1323@gmail.com

The equity market indices started to drop in the initial days of 2020 due to the COVID-19 outbreak and almost remained at a standstill for some time (Ghasemieh & Kashef, 2023). This led to high losses on the side of the investors. Despite a huge amount of studies and works devoted to stock market prediction and the development of efficient models, there are no specific attempts to construct a stable model during the financial crisis. This model shows that big profits can be made and that investors cannot lose money during such moments as the stock market crashes (Qin & Jiang, 2018).

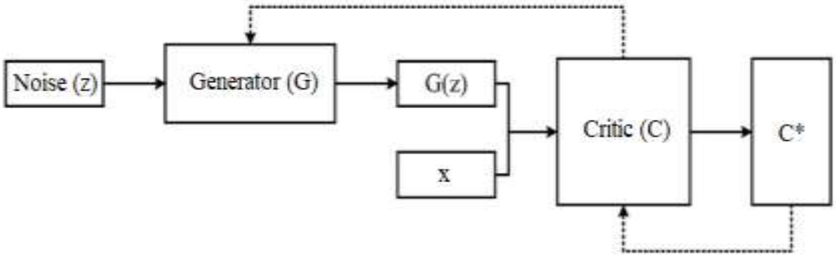

Figure 1: Schematic of WGAN Architecture[9].

1.2. Main Contributions of this research

- Propose the residual connections in the critic network to enhance the gradient flow and the training stability to tackle problems like vanishing gradients.
- Appropriate sampling methods that improved the generation of synthetic data for minority classes and also eliminated class imbalance.
- Applying various noise and sample reshaping techniques which created synthetic data that is as close to real distributions as possible.
- Compare the results to a baseline traditional GAN and IGAN and show that WGAN generated synthetic data with higher quality.
- Stressed the relevance of WGAN for medical diagnosis types of problems where it is critical to get the minority samples right (Hamid et al., 2022)..
- Identifying potential future works in refining the WGAN structures and applying them in different areas of interest [11].

Here we discuss of introduction portion: In section 1 we introduce smartphone history and its usage with different applications, in section 2 we describe on literature review and in section 3 we work on the methodology section, in chapter 4 we discuss outcomes, in section 5 we describe discussion and results and in section 6 we describe conclusion and future work.

2. Methods

In this paper, we focused on proposing a more improved version of the WGAN to achieve the objective of oversampling imbalanced datasets. The initial processes are data loading and data preparation, specification of Generator and Critic architectures, training of WGAN, and assessment of the model based on different performance measures. So, by the transition from the I-GAN to the WGAN, we make use of the Wasserstein distance, which results in a more stable and efficient oversampling method for imbalanced datasets. This approach can enhance the performance of classifiers that have been trained from imbalanced data hence enhancing fairness in machine learning models. The state-of-the-art class balancing method known as the Wasserstein Generative Adversarial Network (WGAN) was introduced to address the class imbalance problem in the model development.

In developing an enhanced Wasserstein Generative Adversarial Network (WGAN) for oversampling imbalanced datasets, three key strategies are implemented: Residual Connections in the Critic Network, Enhanced Sampling for Minority Classes, and Noise and Sample Reshaping. All of these are discussed as follows:

2.1. Residual Connections in the Critic Network

Residual connections are useful in preventing vanishing gradients and allowing for the training of deeper models. In the improved WGAN model, the residual connections are added to the network of the Critic, which aids in better learning capacity of the network and produces a better approximation of the Wasserstein distance (Engelmann & Lessmann, 2021).

2.2. Implementation Steps

- Explaining the basic structure of the Critic network with layers such as convolutional layers, batch normalization, and activation layers (Man, Quddus, Theofilatos, Yu, & Imprialou, 2022).
- Introduce short-cut connections that bypass one or more layers, which means that instead of the normal hierarchical structure, the input to a layer is directly connected to the output of another layer.
- Make the dimensions of the output match for the shortcut connections by either padding or projection layers.
- Train the Critic network using the Wasserstein loss function, which is to minimize the Earth Mover’s Distance (EMD) between the real and the generated samples (Hamid et al. 2023).

2.3. Enhanced Sampling for Minority Classes

To treat class imbalance, a method called enhanced sampling is used, which is aimed at getting more instances of the minority class. This is useful in handling the problem of data imbalance and increasing the efficiency of the classifier.

2.4. Implementation Steps

- Finding the attributes that split the dataset into the minority and the majority classes.
- While performing each iteration of training, focus the method on creating new synthetic samples for the given minority classes.
- Modify the sampling probability to force the generated samples to include larger proportions of samples from the

minority classes.

- Implement these synthetic samples into the training set to have a balanced distribution of the classes.

2.5. Noise and Sample Reshaping

Noise and sample reshaping are then incorporated to improve the quality of the generated samples as well as the variance of samples within the dataset. This entails feeding the input of the Generator some controlled noise and readjusting the form of the generated samples to resemble as closely as possible the real data distribution. New techniques like the generative adversarial network for speech enhancement have proved to be very efficient when large amounts of data are available, but they are still behind in both the low-data regime and the unseen data learning. The Wasserstein Conditional Generative Adversarial Network-Gradient Penalty speech enhancement system incorporates the elastic network into the objective function to reduce the complexity and enhance the model's performance in the low-resource data condition (Munia, Nourani, & Houari, 2020)..

2.6. Implementation Steps

- Feed noise vectors into the Generator, it is preferable that the noise is different, and spread out across all areas. Use Gaussian noise addition or uniform noise that will force small changes in the input vectors.
- Some precautions can be executed within the Generator to avail sample reshaping techniques to attain the output, which is in correlation with the target distribution.
- It is therefore advised to periodically assess and fine-tune the noise and reshaping parameters for the ensuing samples' quality.

2.7. Oversampling

A simple oversampling technique commonly applied in the computer vision field along with the augmentation of the training set and the combating of overfitting is through the geometric transformation including rotation, cropping, image flipping, and change of colour intensity. Many times, these methods merely produce images that are simple redundant copies of the original data. Furthermore, the high-level features of the geometric transformations do not enhance the data distribution because they only cause an image-level transformation through depth and scale. Hence, the over-sampling technique is required to estimate the data distribution and to create data not for the training set only (Chapaneri & Shah, 2022)..

The most common oversampling scheme is the synthetic minority oversampling technique (SMOTE, where new samples of the minority class are synthesized based on the features between the minority points. It is among the commonly used algorithms in handling imbalanced data sets by synthesizing new samples for the minority class with better performance on the classification. However, it could give high noise when dealing with patterns where the appearance of majority and minor classes is not so well defined (Hamid et al., 2023). That is why such extensions as borderline SMOTE and ADASYN appeared to enhance the classification performance by improving the boundary between the two classes. These oversampling methods might not create more samples that are dissimilar to the actual data at first sight, although they may be different on close examination, particularly where routinely extracting features from the unequal distribution of the data is not easy (Suh, 2021). When using these oversampling techniques, they are unable to decrease the classification bias to the majority class when applied to high-dimensional imbalanced data like images and audio. However, the calculation of the distance used in SMOTE is the Euclidean distance, which is not proper for measuring the dissimilarities among samples in high-dimensional feature spaces. Sometimes, the distance between the target sample and the closest cluster is greater than the distance between the sample and the farthest cluster (Zhang, Wang, Pan, & Pan, 2020).

Recently, generative adversarial networks (GANs) have been presented as a class of generative models that got close to the real data distribution (Zhang, Wang, Pan, & Pan, 2020). Conditional GANs (CGAN) and auxiliary classifier GANs (ACGAN) are modifications of the original GANs in which the training process depends on the labels for the classifier (Li, Chen, Shen, Yang, & Zhu, 2019).

Oversampling has also played a significant role when it comes to enhancing the accurate classification of the minority class, which is usually a big problem in biomedical datasets with imbalanced class distribution (Guan, Zhao, Xue, & Pan, 2024). The main endeavor is to introduce a single-channel biosignal data generation method that builds upon the recent progress in the prevalent image processing-based GANs (Zhang, Duan, Hong, Liu, & Zhang, 2021). It has used a Wasserstein GAN (WGAN) for synthesizing synthetic electrocardiogram (ECG) signals because of their stability during training as well as the fact that the loss function of WGAN relates to the quality of the generated image (Wang, Wang, Cui, & Li, 2020).

Imbalance in the class also adds some degree of difficulty to the ability of classification models to accurately predict results. Some of the common strategies include oversampling the minority class by producing artificial cases (Hamid et al., 2022). Some of the standard machine learning algorithms have issues with such imbalances as they try to maximize the overall accuracy at the expense of the minority class. This can entice models with high precision and recall on the benign class and low precision and recall on the malignant class, meaning possible hazardous misdiagnoses. The principal difficulties in handling imbalanced data include wastage of highly representative features and unrealistic estimation of risk that stems from the lack of minority class data, inadequate representation of the majority class data, and possibly, skewing the decision threshold values (Hamid, Iqbal, Arif, Mahmood, Khan, Kama, et al., 2022).

In this study, we have leveraged an improved WGAN to mitigate the problem of imbalance. WGAN can be used to balance the current dataset since by generating synthetic data that simulates the minority class, the model is provided with a more balanced dataset. In addition to the fact that the features of a given minority class are incorporated into the model and learned from, the general training performance and resilience of such a model are also boosted (Hamid, Iqbal, Abbas, Arif, Brezulianu, & Geman, 2022).

3. Results

The main focus is to convert IGAN into WGAN. For this purpose, the main streamline methodology has been used after which the results have occurred. Some main steps of result generation are included as follows:

3.1. Load and Preprocess the Data

The initial process of improving the WGAN's performance is the loading and preprocessing of the data set. This process starts by loading the data and pre-processing it before feeding it into the selected framework compatible with libraries such as TensorFlow or PyTorch. Data preprocessing is the next step in which normalization is done, missing values are dealt with and the raw data is split into training and validation sets. Particular focus is made on the class imbalance by defining minority and majority classes and then introducing methods for oversampling the minority classes to get a similar number of instances in both classes which is suitable for training.

Table 1: PCA Features.

PCA Feature 1	PCA Feature 2
9.1928	1.9486
2.3878	-3.7682
5.7339	-1.0752
1.2562	-1.9023
10.3748	1.6720
-5.4752	-0.6706

In Table 1: we show that PCA Features 1 values and PCA Features 2 values. İn it we let six iterations values.

3.2. Define the Generator and Critic Models for WGAN

After the dataset is prepared, the architecture of the WGAN must be designed and the Generator and Critic must be defined. The Generator's objective is to produce data samples that would have the same distribution as the real data and the Critic's function is to assess the distributions of real and generated data using the Wasserstein distance. The two models are built using neural networks and define layers, activation functions, and loss functions to achieve the best results. This step is essential to establish the adversarial learning process where the Generator has the task of generating samples that can effectively fool the Critic, at the same time, the Critic develops its capability to distinguish between real and fake data.

3.3. Train the WGAN

Training the WGAN is done iteratively to fine-tune everything ranging from the Generator to the Critic. The training process starts with the Critic network that is aimed at estimating the Wasserstein distance between real and generated samples. The Critic's parameters are updated based on the distance metric for change. At the same time, the Generator changes its parameters to generate deception samples more realistic to fool the Critic. The process of adversarial training is iteratively performed in this manner until convergence, while somehow adjusting the hyperparameters such as learning rates and batch sizes for stability and effectiveness.

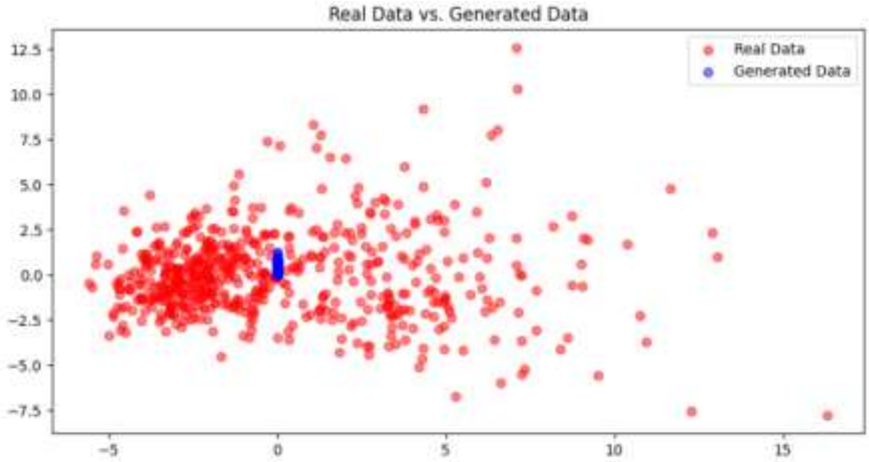

Figure 2: Training Results

Figure 2 result plot shows the difference between the real data and generated data. Real data are denoted by the red dots while the generated data are denoted by the blue dots. Both the axes x and y show the scale for measuring the ratio.

3.4. Evaluate the Model Using Various Metrics

After the training process, the evaluation of the WGAN model is necessary to determine the quality of the synthetic samples. There are usually measures like F1 score, Precision, Recall, and Inception Score used in the determination of quality and the level of diversity in the generated samples of the real data. These measures help in evaluating the level of imbalance in the classes and the general capability of the model in generating improved data for use in the training of machine learning algorithms.

Table 2: Score of Metrics

Metric	Value
AUC Score	0.8169852072211582
F1 Score	0.3941747572815534
Confusion Matrix	[[311, 615], [9, 203]]

In **Table 2** we use different metrics like AUC score, F1 score, and Confusion Metrix with their assigned values which are used in this paper. Value of AUC score = 0.8169852072211582. F1 score = 0.3941747572815534 and confusion matrix = 311, 615], [9, 203].

Table 3: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.97	0.34	0.50	926
1	0.25	0.96	0.39	212
Accuracy			0.45	1138
Macro Avg	0.61	0.65	0.45	1138
Weighted Avg	0.84	0.45	0.48	1138

In Table 3 we generate a classification report in which we set some classes (0, 1, Accuracy, Macro Average, and Weighted Average). We apply different measures in the class which are precision, recall, F1-score, and Support. And find different values according to these classes.

ROC curves and AUC scores are perfect tools for classifier assessment, and the same applies to WGAN models. The ROC curve represents the TPR (sensitivity) and FPR (1-specificity) at different decision points to establish the model's performance. It is an overall measure of the performance of the model in terms of separating the classes and is the mean of the ROC curve. These metrics taken together provide a broad picture of the WGAN's performance and bear out its ability to effectively generate synthetic data that improves the training of machine learning models by addressing class imbalance and boosting classification rates.

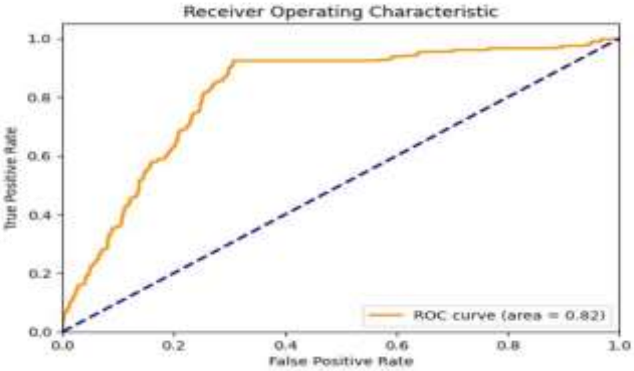

Figure 3: ROC Curve (1)

In Figure 3 AUC-ROC represents a general measurement of the capacity of the given model to effectively distinguish between the positive and the negative results with a value of 0.5 indicating no discrimination and 1.0 indicating perfect classification. An AUC of 0.82 suggests that the model performs well, with a high probability of correctly distinguishing between positive and negative instances. The ROC Curve resides on two axes axis-x and axis-y. On axis x, the false positive ratio is shown while on axis y, the true positive rate is shown.

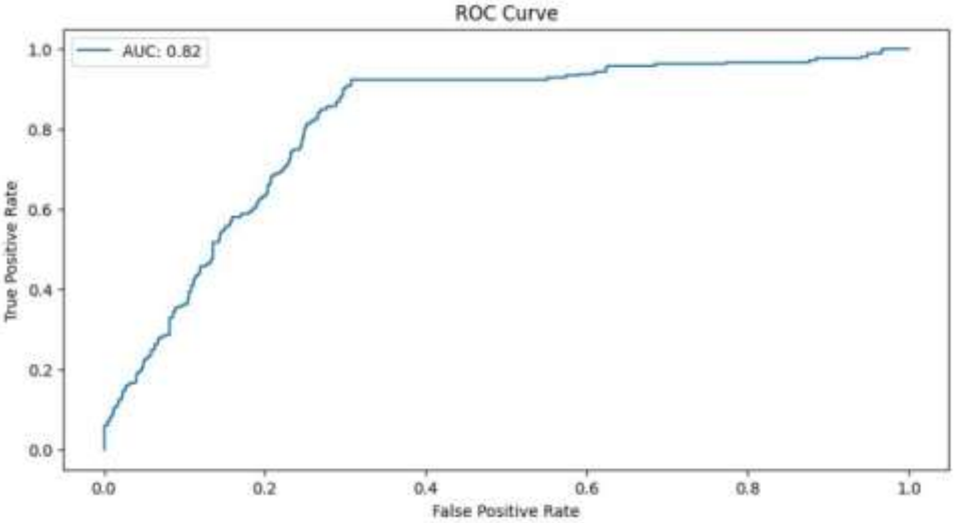

Figure 4: ROC Curve (2).

In Figure 4 ROC Curve, as illustrated with the AUC value of approximately 0. 82, indicates that the proposed model has a fairly acceptable level of classification performance with the area under the curve being greater than 0.5, which means the classification is not performing a random assignment of data to the two classes. Regarding the classification results, the confusion matrix shows that the model is very good at predicting class 0, having high precision and recall for class 0, however, for class 1 the model is extremely poor, with low precision but high recall. The F1 score amounts to 0. 39 represents the balance between the precision and the recall for the minority class, which, in a sense, means that it is a good model that recognizes most of the positive cases, albeit with the correspondingly high number of FPs. Using the classification report, it is noted that the overall accuracy of the classifier is 45 percent, class 0 has slightly better accuracy compared to class 1, and overall, the macro average F1 score is 0. 45 reveals significant differences in performance by classes.

3.5. Comparison to IGAN Model

3.5.1. IGAN Model Limitations

3.5.1.1. No Residual Connections

The original IGAN model does not have residual connections within its structure, which causes the issue of gradient vanishing and a drop in the training process stability. Executing connections is core to deep neural networks as they allow gradients to flow through the network backpropagation process to achieve better learning and convergence.

3.5.1.2. Class Imbalance

IGAN does not contain specific procedures for dealing with class imbalance in the datasets. This is evident mainly when there is a variation of classes in the dataset where the generator provides a poor representation of the classes during the synthetic data creation.

3.5.1.3. Unconstrained Generation

When IGAN lacks sample reshaping methods, it may synthesize samples far from the desired distribution. In particular, this unrestricted generation process can lead to the generation of synthetic data of lower quality, which hinders their application in tasks such as training models in machine learning.

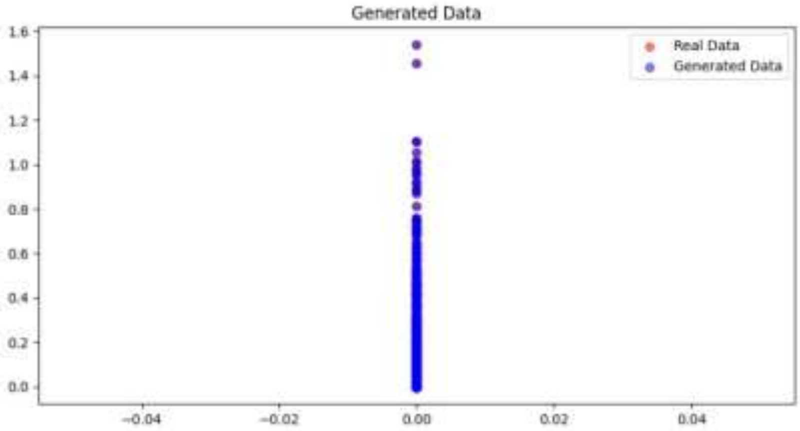

Figure 5: Generated Data

In Figure 5 the scatter plot of the generated data by the generator is to elucidate how effectively the model can generate data in different chunks that might be similar to the real data.

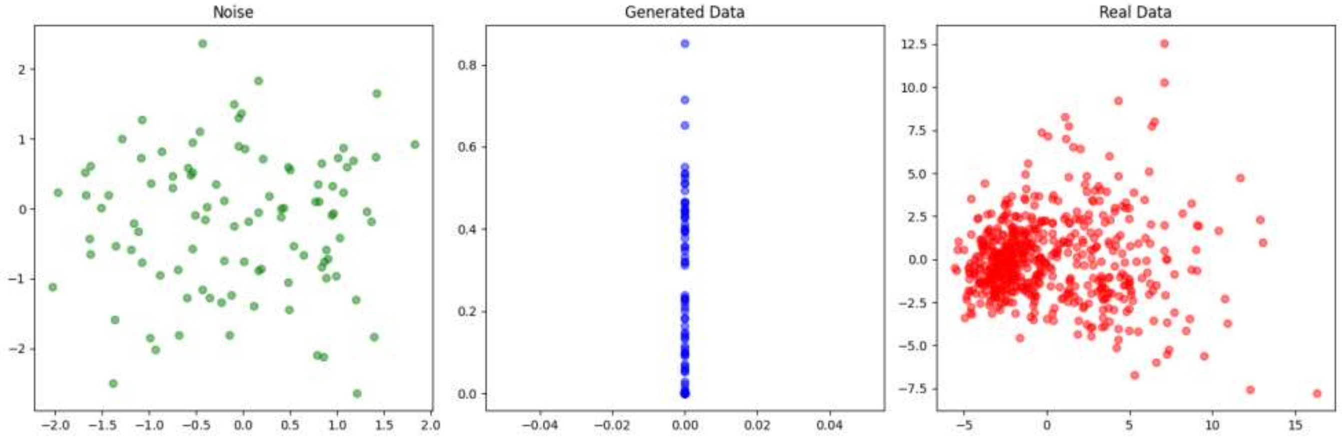

Figure 6: Comparison Graphs

The above visualize_pipeline (Figure 6) function creates a three-panel plot to compare one or more stages or layers in a Generative Adversarial Network (GAN). The first panel represents a scatter plot of noise input which is usual and mainly used to produce synthetic data. The second one is the scatter plot of the generated data by the generator to elucidate how effectively the model can generate data that might be similar to the real data. The third panel shows the distributions of the actual data to which the generated data is trying to fit as a benchmark. This function allows inspecting the data transformation from noise toward the generated samples and it assists in quantifying GAN's output.

3.6. Enhanced WGAN Model Advantages

3.6.1. Enhanced Stability and Learning

The IGAN model's stability issues are resolved by the more advanced Wasserstein Generative Adversarial Network (WGAN) with the inclusion of residual connections in the critic network. These connections add better gradient flow and bring about more stable and faster training processes, so helping the model learn from the features of data suitably.

In data mining, standard classification techniques can no longer learn from big imbalanced data sets. Oversampling solves this issue by generating data for the minority class so that the two classes are balanced before feeding them to the model. The Traditional oversampling approaches are based on the Synthetic Minority Oversampling Technique (SMOTE), which is a technique relying on the information of the local surroundings and results in insufficiently realistic data. While the Generative Adversarial Network (GAN) aims to model the true distribution to generate data for the minority class. However, both of them are challenging due to the issues of mode collapse and unstable training. These problems getting resolved by using WGAN.

3.6.2. Balanced Class Distribution

The class imbalance problem is solved in Enhanced WGAN by using expert sampling techniques on minority classes. This way, the generator receives enough exposure to all classes and therefore the generation of synthetic data is more balanced.

3.6.3. Controlled and Realistic Samples

The addition of the reshaping function in the EWGAN restricts the output samples and makes them closer to the target distribution. This controlled generation process leads to realistic and high quality of the synthetic data which makes the supplementation of imbalanced datasets more efficient for the development of machine learning models with better robustness and accuracy.

3.6.4. Confusion Matrix

Confusion matrices which are also called error matrices are widely used when comparing the results of models of classification. They enable the determination of how accurately a model categorizes the instances into various classes, about their actual and predicted status. In most cases, confusion matrices are two-dimensional matrices whose rows indicate the instances in a particular predicted class and whose columns represent the instances in a particular actual class.

When it comes to the confusion matrix, the main diagonal elements represent the correct classification of instances by the model for the respective class while off diagonal element represents the misclassification by the model. In particular, the entry of the matrix for the ith row and the jth column shows the number of instances that belong to the ith class, yet the model predicts it to be in the jth class. This layout makes it possible to find out such performance indicators as accuracy, precision, recall, and F1-score, which are obtained from different values of the matrix.

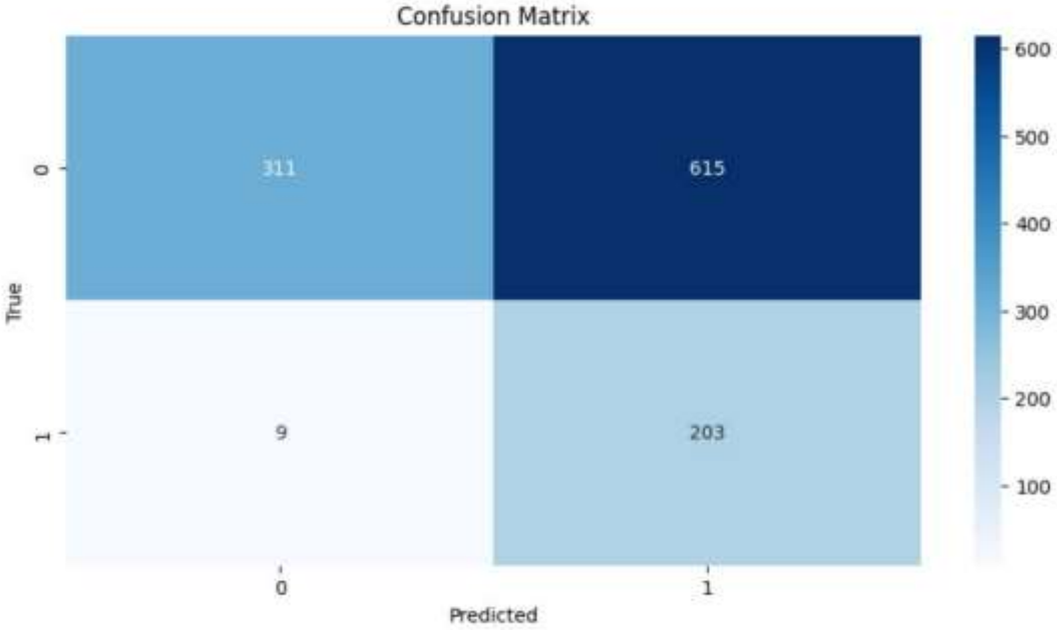

Figure 7: Confusion Matrix

In Figure 7 Confusion matrices provide more details about a model's performance by providing instances of misclassification. For example, they can differentiate whether the model has a higher tendency to misclassify particular classes or whether the model is accurate across the classes. This level of examination is especially useful in cases of uneven datasets, where even such basic indicators as accuracy may not always be an accurate reflection of the model's performance. In addition to providing a visual and numeric summary of classification results, confusion matrices offer practical information about modifying model parameters, features, or samples to improve classification accuracy and dependability.

3.7. Final Histogram View

Histogram charts collectively represent the distribution of data in each of the columns in the dataset and therefore are of great importance when it comes to analyzing the characteristics of the data and the need for some preprocessing. They all show the distribution of values within one feature, establishing the nature, whether skewed, normal, or of some other sort. This visual representation helps in finding out the outliers, observing the distribution of values, and deciding whether normalization or scaling should be applied to have a uniform distribution of features.

In Figure 8 histograms, analysts can easily understand the center and spread of each feature which is important in the subsequent modeling processes. For example, attributes with highly skewed distribution may be normalized through a logarithmic transformation because such variables tend to improve the performance of models. On the other hand, some features might have a normal distribution, and usually, do not need much preprocessing as their raw form is suitable for modelling tasks. Overall, these classes of histogram charts are the initial examination tool to facilitate data preprocessing and determine the distribution of the collected data for subsequent analyses and modelling with more complex methods.

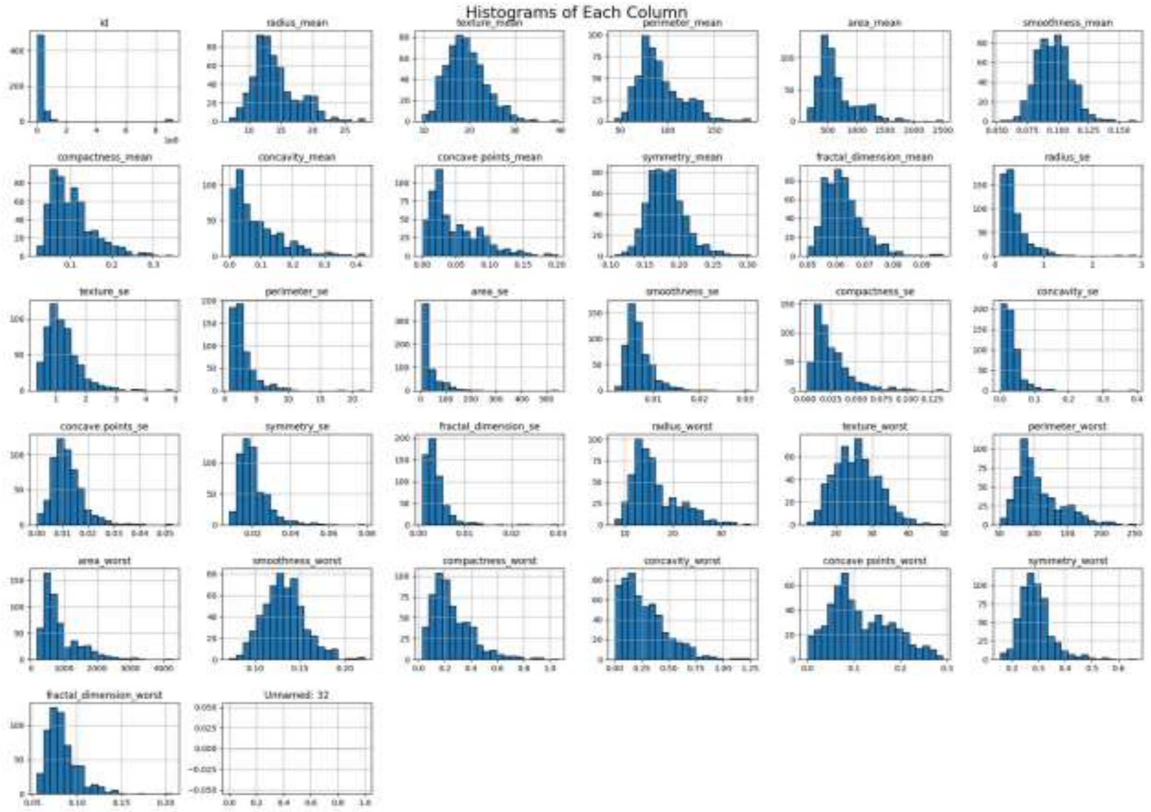

Figure 8: Histogram for Each Column

Table 4: Comparison Table for each column

Feature	Description	Range/Values	Observations
texture_worst	The texture of the worst area of the tumour	17.33 - 26.50	The values vary significantly, indicating different textures in the worst tumour area.
perimeter_worst	The perimeter of the worst area of the tumour	98.87 - 184.60	Shows a wide range, reflecting the variability in tumour size.
area_worst	Area of the worst area of the tumour	567.7 - 2019.0	The area values are distributed over a broad spectrum, with some outliers.
smoothness_worst	The smoothness of the worst area of the tumour	0.1238 - 0.2098	The values are relatively close, indicating moderate variability in smoothness.
compactness_worst	Compactness of the worst area of the tumour	0.1866 - 0.8663	Shows significant variation, highlighting differences in tumour compactness.
concavity_worst	Concavity of the worst area of the tumour	0.2416 - 0.7119	The values indicate varying degrees of concavity.
concave points_worst	Number of concave points in the worst area of the tumour	0.1625 - 0.2654	Displays a moderate range, suggesting variability in the number of concave points.
symmetry_worst	Symmetry of the worst area of the tumor	0.2364 - 0.6638	Symmetry values show considerable variation, affecting the appearance of tumour symmetry.
fractal_dimension_worst	The fractal dimension of the worst area of the tumour	0.07678 - 0.17300	The fractal dimension values vary, indicating different levels of complexity in tumour structure.
Unnamed: 32	Possibly an identifier or irrelevant feature	NaN	The presence of NaN values suggests that this feature may not be useful or has missing data.

In Table 4 we show a comparison of different features in which we discuss their features, Range, and observations. and find out the best one for further usage in this paper.

3.8. Key Observations

- **Wide Range of Values**

Features such as perimeter, area, and fractal dimension exhibit a wide range, essential for distinguishing between different tumour characteristics.

- **Variability**

There is significant variability in texture, compactness, and symmetry, which are crucial for diagnostic analysis.

- **Potential Outliers**

Some features, particularly area and perimeter, show potential outliers, which may need further investigation or preprocessing.

4. Discussion

The data set that is used in this study is the Breast Cancer Wisconsin (Diagnostic) Data Set for training and testing of the proposed WGAN model. This dataset contains quantitative characteristics of the cell nuclei present in the fine needle aspirate images of breast masses obtained through digitization Hence this dataset has features including but not limited to radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensions. In each of them, the case is attributed as either benign or malignant, so it is quite suitable for binary classification tasks in the context of breast tumour diagnosis. The database is composed of 569 instances; 357 of them are classified as benign, while 212 are classified as malignant, which signifies a severe case of the class imbalance problem that goes against the traditional machine learning algorithms. With the help of WGAN, it planned to obtain highly realistic synthetic samples to address the problem of class imbalance and improve the performance of the predictive models on the minority class. This approach is particularly significant in early diagnosis of the disease since the prognosis is accurate and reliable hence improving the probabilities of the breast cancer disease.

The dataset comprises essential tumour characteristics crucial for medical diagnosis and classification. Texture, Perimeter, and Area (Worst) exhibit substantial variability, indicating diverse tumour sizes and textures across samples. Smoothness, Compactness, and Concavity (Worst) metrics reveal significant irregularities and growth patterns, pivotal for assessing tumour malignancy. Symmetry and Fractal Dimension (Worst) provide insights into structural uniformity and complexity, influencing diagnostic decisions. However, Unnamed: 32 contains NaN values, suggesting data issues that may require resolution to ensure comprehensive analysis and model efficacy.

The conventional method of signal processing for fault detection involves using fault diagnosis algorithms to identify the specific fault from the raw vibrational signal. Traditional FD techniques using the time and frequency domain features including wavelet transform, variational mode decomposition, and permutation entropy have already been employed in previous studies to enhance the fault diagnosis capability. Besides, intelligent fault diagnosis methods have been successfully developed in recent years. Especially, fault diagnosis approaches based on data include its characteristic, deep learning (DL) and the method offers a better solution to fault classification since there is no need for prior knowledge of the automatic system. For DL-based methods, the most important precondition is the availability of large amounts of balanced data to sufficiently train deep neural networks. But in most time of industrial processes, rolling bearing is often in normal working conditions and fault signals are difficult to obtain. It is noticeable that in a normal state, the acquired data amount is far more than that in a faulty state. The proportion of normal samples and fault samples is not rational, which also results in a significant decline in the performance of fault diagnosis.

4.1. Enhanced WGAN Strategies: Findings and Benefits

4.1.1. Residual Connections in Critic Network

The integration of residual connections improves gradient flow as well as the stability of the training process. This optimization helps in faster convergence and enhances the model's ability to handle difficult datasets such as medical imaging, which is vital. Better gradient flow helps avoid problems such as gradient vanishing, providing stable training for the model and its further improvement.

4.1.2. Enhanced Sampling for Minority Classes

Enhancing sampling techniques for the minority class improves the representation of such classes and diversifies the training data. This strategy helps to lessen class imbalance which is extremely important when creating synthetic samples with proper distribution of classes. This way, the model can increase the representation of the minority classes during synthetic data generation and improve the overall model and its corresponding, while at the same time decreasing the bias.

Imbalanced data classification is a natural phenomenon that occurs frequently in the fields of machine learning and data mining. While imbalanced data classification can be broadly defined as the distribution of samples in different categories of a data set, the term is more specifically used to describe the condition when one or more categories contain significantly fewer samples than others. In any binary or multiclass dataset having the problem of imbalanced data classification, the group that contains a small number of samples is called a positive or minority class, while the one with a large number of samples is termed a negative or majority class. In real-world scenarios, imbalanced data classification is linked to text classification, consumer's buying behaviour from the seen behavioural history, credit card fraud detection from the transaction data, and diagnosis of diseases from medical data. Since the classification algorithms inherited from the traditional methods are fit for cases where the data is distributed in a roughly equal manner, the use of such algorithms when it comes to the classification of imbalanced data may result in varying degrees of deficiencies and so become rather ineffective. So, improving the classifier's performance when working with imbalanced data and increasing the ability to recognize the minority class is an important issue that needs to be solved.

4.2. Noise and Sample Reshaping

The use of filtered noise and redesigning the mechanisms of control enhances the acoustics and authenticity of synthesized samples. This approach helps in making the generated data approximate the target distribution, thus enhancing the usefulness of synthetic data to other tasks. Here, the model transforms samples according to the characteristics acquired during the learning process to generate realistic data that, in terms of subsequent machine learning, is closer to the actual data distribution.

Table 5: Features Analysis

Enhanced WGAN Strategies	Features Addressed
Residual Connections in Critic Network	Gradient flow, and stability during training
Enhanced Sampling for Minority Classes	Minority class representation, class imbalance
Noise and Sample Reshaping	Synthetic data quality, alignment with target distribution

In Table 5 we examine the feature analysis of WGAN (Wasserstein Generative Adversarial Network). Enhance their strategies and Features addressed.

4.3. Comparison of WGAN with IGAN and GAN

In Table 6 we compare GAN (Generative Adversarial Networks), IGAN (Improved Generative Adversarial Networks), and WGAN (Wasserstein Generative Adversarial Network). We compare them based on features like Architecture, Loss Function, Training Stability, etc.

Table 6: Comparative Table for GAN, IGAN, and WGAN.

Feature	GAN	IGAN	WGAN
Architecture	Simple, two neural networks (Generator, Discriminator)	Adds enhancements like deeper networks and regularization	Similar to GAN, but with Wasserstein distance metric
Loss Function	Minimax (Jensen-Shannon divergence)	Minimax with improved training stability	Wasserstein distance improves training stability
Training Stability	Prone to mode collapse, unstable training	Improved with deeper networks, regularization	Enhanced with gradient penalty, stable training
Gradient Flow	Limited, suffers from vanishing gradients	Enhanced with deeper networks	Improved by Wasserstein distance metric, stable
Handling Mode Collapse	Common issue, diversity of generated samples compromised	Mitigates with improved loss functions	Mitigates with Wasserstein distance, diverse samples
Quality of Generated Samples	Variable, prone to producing low-quality samples	Improved, more realistic samples	High quality, realistic samples, diverse outputs
Robustness to Noise	Limited, sensitive to noise in training data	Improved with regularization techniques	Handles noise better, more robust training
Class Imbalance Handling	Not explicitly addressed	Limited improvement	Enhanced through sampling techniques for balanced output
Applications	Various, including image generation, data augmentation	Medical imaging, data synthesis	Medical imaging, data synthesis, image generation
Advantages	Simple architecture, versatile applications	Improved stability, better sample diversity	Stable training, high-quality outputs, robust to noise
Disadvantages	Prone to mode collapse, unstable training	Limited improvement in stability	Requires careful tuning, computational complexity

4.4. Limitations of Research

- **Computational Complexity**

WGAN strategies have been improved by the addition of residual connections, better sampling techniques, and noise reshaping, but these enhancements come with a large computational overhead. This can lead to increased training times and required resources and thus could pose some scalability issues on the proposed methods. In the real-world contexts especially when dealing with big data or streaming data this can impact the viability and application of the improved models.

- **Careful Tuning Required**

As the paper demonstrates, improved WGANs require careful calibration of hyperparameters and architectural settings to obtain the best performance. Unfortunately, this requirement for careful tuning leads to variability that is inherent in the system and may give unequal results in different experiments or data sets. Hence it is seen that if the network’s hyperparameters are not well tuned then the improvements that may be derived may not be optimal affecting the model’s efficiency.

- **Complexity of Noise and Sample Reshaping**

Readapting the samples from the previous stage and reducing noise are also challenges confined to training. These techniques enhance the quality and fitness of synthetic data for the target distributions but at the same time cause extra computational load. These additional intricacies could result in extended training durations and increased computational requirements, which is important in determining the feasibility of implementing these improvements in actual applications.

5. Conclusion

The use of Wasserstein Generative Adversarial Networks (WGAN) has been proven to be quite effective in addressing problems arising from imbalanced data sets, especially in sectors such as medical diagnosis. These results highlight WGAN’s potential to deal with problems related to the vanishing of gradient and improve the sample generation using an innovative architecture and better sampling strategies. This is why thanks to the residual connections, and the sophisticated data generation methods WGAN can generate high-quality synthetic data to balance the classes and improve the training of the model. These findings substantiate that WGAN has a qualitatively better performance than basic GAN and IGAN in synthesizing data that is closer to actual distributions. Enhanced WGAN has been shown to outperform the method by having better values in the AUC and ROC curves, which indicates the extent of addressing the class imbalance challenge and improved classification.

The results also show that there is still room for improvement, in terms of more accurate results in some classes. Directions for future work include refining the WGAN architectures, examining the potential of WGANs in multiple domains, especially healthcare, and creating approaches for improving the synthesis of realistic and diverse datasets. Such development can contribute to enhancing the methods of machine learning to improve the handling of imbalanced datasets and provide far superior and superior predictive models.

References

Ali, G., Dastgir, A., Iqbal, M. W., Anwar, M., & Faheem, M. (2023). A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. *IEEE Journal of Translational Engineering in Health and Medicine*, *11*, 341-350.

Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214-223). PMLR.

Chapaneri, R., & Shah, S. (2022). Enhanced detection of imbalanced malicious network traffic with regularized generative adversarial networks. *Journal of Network and Computer Applications*, *202*, 103368.

Dewi, C., Chen, R. C., & Liu, Y. T. (2021). Wasserstein generative adversarial networks for realistic traffic sign image generation. In *Intelligent Information and Database Systems: 13th Asian Conference, ACIIDS 2021, Phuket, Thailand, April 7–10, 2021, Proceedings 13* (pp. 479-493). Springer International Publishing.

Engelmann, J., & Lessmann, S. (2021). Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, *174*, 114582.

Ghasemieh, A., & Kashef, R. (2023). An enhanced Wasserstein generative adversarial network with Iranian angular fields for efficient stock market prediction during market crash periods. *Applied Intelligence*, *53*(23), 28479-28500.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Guan, S., Zhao, X., Xue, Y., & Pan, H. (2024). AWGAN: An adaptive weighting GAN approach for oversampling imbalanced datasets. *Information Sciences*, *663*, 120311.

Hamıd, k., muhammad, h. a. b., ıqbal, m. w., hamza, m. a., bhattı, s. u., Hassan, s. a., & ıkram, a. extendable banhattı sombor ındıces for modelıng certaın computer networks.

Hamid, K., Iqbal, M. W., Arif, E., Mahmood, Y., Khan, A. S., Kama, N., ... & Ikram, A. (2022). K-Banhatti Invariants Empowered Topological Investigation of Bridge Networks. *Computers, Materials & Continua*, *73*(3).

Hamid, K., Iqbal, M. W., Ashraf, M. U., Gardezi, A. A., Ahmad, S., Alqahtani, M., & Shafiq, M. (2023). Intelligent Systems and Photovoltaic Cells Empowered Topologically by Sudoku Networks. *Computers, Materials & Continua*, *74*(2).

Hamid, K., Iqbal, M. W., Virk, A. U. R., Ashraf, M. U., Alghamdi, A. M., Bahaddad, A. A., & Almarhabi, K. A. (2022). K-Banhatti Sombor Invariants of Certain Computer Networks. *Computers, Materials & Continua*, *73*(1).

Hamid, K., Waseem Iqbal, M., Abbas, Q., Arif, M., Brezulianu, A., & Geman, O. (2022). Discovering irregularities from computer networks by topological mapping. *Applied Sciences*, *12*(23), 12051.

Jin, Q., Lin, R., & Yang, F. (2019). E-WACGAN: Enhanced generative model of signaling data based on WGAN-GP and ACGAN. *IEEE Systems Journal*, *14*(3), 3289-3300.

Lee, G. C., Li, J. H., & Li, Z. Y. (2023). A Wasserstein Generative Adversarial Network–Gradient Penalty-Based Model with Imbalanced Data Enhancement for Network Intrusion Detection. *Applied Sciences*, *13*(14), 8132.

Li, Q., Chen, L., Shen, C., Yang, B., & Zhu, Z. (2019). Enhanced generative adversarial networks for fault diagnosis of rotating machinery with imbalanced data. *Measurement Science and Technology*, *30*(11), 115005.

Man, C. K., Quddus, M., Theofilatos, A., Yu, R., & Imprialou, M. (2022). Wasserstein generative adversarial network to address the imbalanced data problem in real-time crash risk prediction. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 23002-23013.

Munia, M. S., Nourani, M., & Houari, S. (2020, November). Biosignal oversampling using Wasserstein generative adversarial network. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 1-7). IEEE.

Qin, S., & Jiang, T. (2018). Improved Wasserstein conditional generative adversarial network speech enhancement. *EURASIP Journal on Wireless Communications and Networking*, *2018*(1), 181.

Suh, S. (2021). *Improving Classification Performance under Imbalanced Data Conditions using Generative Adversarial Networks* (Doctoral dissertation, Technische Universität Kaiserslautern).

Suh, S., Lee, H., Lukowicz, P., & Lee, Y. O. (2021). CEGAN: Classification Enhancement Generative Adversarial Networks for unraveling data imbalance problems. *Neural Networks*, *133*, 69-86.

Wang, W., Wang, C., Cui, T., & Li, Y. (2020). Study of restrained network structures for wasserstein generative adversarial networks (WGANs) on numeric data augmentation. *IEEE Access*, *8*, 89812-89821.

Zhang, H., Wang, R., Pan, R., & Pan, H. (2020). Imbalanced fault diagnosis of rolling bearing using enhanced generative adversarial networks. *IEEE Access*, *8*, 185950-185963.

Zhang, L., Duan, L., Hong, X., Liu, X., & Zhang, X. (2021). Imbalanced data enhancement method based on improved DCGAN and its application. *Journal of Intelligent & Fuzzy Systems*, *41*(2), 3485-3498.

Zhang, M., Liu, Y., Luan, H., & Sun, M. (2017, September). Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1934-1945).

Zheng, M., Li, T., Zhu, R., Tang, Y., Tang, M., Lin, L., & Ma, Z. (2020). Conditional Wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification. *Information Sciences*, *512*, 1009-1023.