Synthetic generation of multidimensional data to improve classification model validity

Al-Qerem, Ahmad; Ali, Ali Mohd; Attar, Hani; Nashwan, Shadi; Qi, Lianyong; Moghimi, Mohammad Kazem; Solyman, Ahmed

Published in:
Journal of Data and Information Quality

DOI:
10.1145/3603715

Publication date:
2023

Document Version
Author accepted manuscript

Synthetic Generation of Multidimensional Data to Improve Classification Model Validity

AHMAD AL–QEREM, Department of Computer Science, Zarqa University, Zarqa, Jordan
ALI MOHD ALI, Communications and Computer Engineering Department, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan
HANI ATTAR, Department of Energy Engineering, Zarqa University, Zarqa, Jordan
SHADI NASHWAN, College of Computer and Information Sciences, Jouf University, Aljouf, Saudi Arabia
LIANYONG QI, Department of Computer Science, China University of Petroleum (East China), China
MOHAMMAD KAZEM MOGHIMI, Department of Communications Engineering, University of Sistan and Baluchestan, Zahedan, Iran
AHMED SOLYMAN, Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Nişantaşı University, İstanbul, Turkey

This article aims to compare Generative Adversarial Network (GAN) models and feature selection methods for generating synthetic data in order to improve the validity of a classification model. The synthetic data generation technique involves generating new data samples from existing data to increase the diversity of the data and help the model generalize better. The multidimensional aspect of the data refers to the fact that it can have multiple features or variables that describe it. The GAN models have proven to be effective in preserving the statistical properties of the original data. However, the order of data augmentation and feature selection is crucial to build robust and accurate predictive models. By comparing the different GAN models with feature selection methods on multidimensional datasets, this article aims to determine the best combination to support the validity of a classification model in multidimensional data.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**; • **Security and privacy** → **Intrusion detection systems**; • **Networks** → **Sensor networks**;

Additional Key Words and Phrases: Multidimensional data, model validity, data augmentation, filter method, wrapper method

1 INTRODUCTION

Machine learning algorithms have shown great promise in the field of medicine [1–3], and the increasing use of electronic health records (EHRs) has made the use of these algorithms more feasible [4]. However, the high-dimensional and complex nature of medical data presents many challenges that can limit the accuracy and validity of these models. One such challenge is the limited size and diversity of the available data, which can result in bias and reduce the generalizability of the models [5]. In addition, many datasets have missing values or inconsistent data, which can lead to imprecise predictions and misclassification of patients [6, 7].

To address these limitations, synthetic generation of multidimensional data has become an important tool in medical research [8]. By creating a large and diverse dataset that is representative of the real-world population, researchers can improve the validity of the models and reduce the risk of bias. Furthermore, synthetic generation can also be used to address missing data by imputing values in a statistically valid manner. One popular method of synthetic data generation is the use of generative adversarial networks (GANs) [9, 10]. GANs are deep learning models that use two neural networks to generate synthetic data that is indistinguishable from real data. The first network generates synthetic data, whereas the second network acts as a discriminator that assesses whether the data are real or synthetic. The two networks are trained in an adversarial manner, with the generator learning from the feedback of the discriminator.

Another method for synthetic data generation is the use of simulation models [11]. These models use mathematical equations to simulate the behavior of a real-world system, such as a patient's health over time. By using a simulation model, researchers can generate large amounts of synthetic data that are representative of the real-world population, while also controlling for important variables such as age, gender, and disease state. Multimodality and multidimensionality are important characteristics of modern data [12, 13]. *Multimodality* refers to the presence of multiple sources of information within a single dataset [14]. For example, a dataset could contain both numerical and categorical variables, images, audio, and text. The combination of these sources provides a more complete picture of the data, allowing for more informed decisions to be made based on the data. *Multidimensionality* refers to the presence of multiple features or variables that describe the data [15]. This results in a more complex representation of the data, as there are many variables that can impact the outcome being studied. For example, a dataset of patients with a certain disease might contain variables such as age, gender, family history, and previous medical conditions. Each of these variables can play a role in the outcome, making it important to consider all of the variables when analyzing the data. The presence of multimodality and multidimensionality in modern data provides a more complete and complex representation of the data. This enables more informed decisions to be made and leads to a better understanding of the underlying relationships between the variables in the data [16, 17].

Augmented multidimensional data refers to additional data that have been generated or derived from the original data to enhance its representational power. The goal of augmenting the data is to increase its diversity and to provide more information for the machine learning models to learn from. *Multidimensional data* refers to data that have multiple features or variables that describe it. By augmenting the data, more information can be provided to the machine learning models,

Table 1. Description of Data In Terms of Multimodality and Multidimensionality

Data Type	Multimodality	Multidimensionality
Image Data	High	High
Audio Data	High	High
Text Data	Low	High
Clinical Data	Medium	Medium
Genomic Data	High	High
Video Data	High	High

which can improve their performance and accuracy. Augmented data can be created using various techniques, such as data augmentation, synthetic data generation [18], and feature engineering [19]. These techniques can be used to create more diverse data, to improve the representation of the original data, and to help the machine learning models generalize better to new data. Augmented multidimensional data is an important aspect of machine learning and is widely used in various applications, such as image recognition, speech recognition, and natural language processing. Table 1 presents a description of data in terms of multimodality and multidimensionality [20]. It contains various types of data and their levels of multimodality and multidimensionality. Image data is considered highly multimodal as it contains multiple sources of information and is highly multidimensional as it contains information about multiple aspects of the image [21, 22]. Audio data is also considered highly multimodal and multidimensional for similar reasons [23]. Text data is low in multimodality but high in multidimensionality as it contains information about multiple aspects of the text. Clinical data is medium in multimodality and multidimensionality, whereas genomic data is considered highly multimodal and multidimensional due to its genetic information content. Video data is also highly multimodal and multidimensional as it contains moving images, audio, and context information [24].

Synthetic data generation has become an increasingly popular technique in fields such as machine learning, data science, and data privacy. However, there are several potential limitations to consider when using synthetic data. One limitation is the potential for bias and lack of representativeness, as synthetic data heavily relies on the quality and quantity of the original data used for the model training. Another limitation is the potential loss of information, as the synthetic data may not accurately capture all the information present in the original data, leading to potential errors in downstream analysis. Privacy and security concerns are also a limitation, as attackers can use synthetic data to infer sensitive information about individuals. Scalability and flexibility are other potential limitations, as generating synthetic data for large datasets may require significant computing power and time, and the generated synthetic data may not be easily adaptable to new situations. Finally, legal and ethical issues may also arise, particularly for applications such as medical or financial, where synthetic data may not be an accurate representation of real-world situations. In conclusion, while synthetic data generation offers many potential benefits, careful consideration should be given to its limitations and potential risks.

2 SIOT-BASED HEALTHCARE SERVICES

The advancement of technology has enabled the collection of health data through the integration of the social Internet of Things (SIoT) and smart apps. These technologies offer a convenient and effective way to monitor and track an individual's health status, and can provide valuable insights for healthcare providers. In this section, we outline the steps involved in collecting, updating, storing, and processing health data through the SIoT and smart apps.

Table 2. SIoT-Based Healthcare Data in Terms of Multimodality and Multidimensionality

Data Type	Multimodality	Multidimensionality
Vital Sign Monitoring	Continuous Stream of Data	Multiple Dimensions: Heart Rate, Blood Pressure, Respiratory Rate, Oxygen Saturation, etc.
Physical Activity Monitoring	Continuous Stream of Data	Multiple Dimensions: Steps Taken, Distance Traveled, Calories Burned, etc.
Sleep Monitoring	Continuous Stream of Data	Multiple Dimensions: Sleep Duration, REM Sleep, Deep Sleep, etc.
Medication Adherence Monitoring	Discrete Event Data	Single Dimension: Time of Medication Intake
Nutrition Tracking	Discrete Event Data	Multiple Dimensions: Meal Type, Caloric Intake, Nutrient Composition, etc.

Data Collection: Health data can be collected through various sources such as wearable devices, smart health apps, and personal medical records. Wearable devices such as fitness trackers, smartwatches, and glucose monitors can continuously monitor an individual's health status and provide real-time health data. Smart health apps allow individuals to manually input data such as symptoms, medication intake, and food consumption, and can also integrate data from wearable devices.

Data Updating: Health data collected through the SIoT and smart apps needs to be updated regularly to maintain its relevance and accuracy. Wearable devices can automatically update health data in real time, whereas manual inputs through smart health apps can be updated as often as desired.

Data Storage: Health data collected through the SIoT and smart apps needs to be securely stored to ensure privacy and prevent unauthorized access. Cloud-based storage solutions, such as Amazon Web Services or Microsoft Azure, provide a scalable and secure storage option for health data.

Data Processing: Health data collected through the SIoT and smart apps can be processed using various techniques such as data mining, machine learning, and artificial intelligence [24]. Data processing can provide valuable insights into an individual's health status and help healthcare providers make informed decisions. The integration of the SIoT and smart apps in healthcare provides a convenient and effective way to collect, store, and process health data. These technologies have the potential to revolutionize the healthcare industry and improve patient outcomes. Table 2 contains a description of SIoT-based healthcare data in terms of multimodality and multidimensionality.

In Table 2, we present a few examples of the data that can be collected through the SIoT and smart apps for healthcare services. The data is categorized based on whether it is a continuous stream of data or discrete event data. The multimodality of the data refers to whether it is collected in a continuous or discrete manner, whereas the multidimensionality of the data refers to the number of different aspects of the data that are being collected. For example, vital sign monitoring data is a continuous stream of data that is collected over time, and it includes multiple dimensions such as heart rate, blood pressure, respiratory rate, and oxygen saturation. On the other hand, medication adherence monitoring data is discrete event data that only includes the time of medication intake.

3 MATERIAL AND METHOD

In the field of machine learning and data analysis, the order of data augmentation and feature selection is a crucial aspect of the methodology. The goal is to build robust and accurate predictive

Table 3. Properties of Experimental Datasets

Dataset	# Instances	# Features	# Classes	Missing Values	Task Type	Modality
Appendicitis	106	7	2	No	Binary	Medical
Australian	690	14	2	Yes	Binary	Financial
Haberman	306	3	2	No	Binary	Medical
Ionosphere	351	34	2	No	Binary	Physical
Liver	345	6	2	Yes	Binary	Medical
Parkinsons	195	23	2	No	Binary	Medical
Phoneme	5,140	5	2	No	Binary	Audio
WDBC	569	30	2	No	Binary	Medical

models that can be applied to real-world problems. The answer to which method is better, feature selection or data augmentation, largely depends on the problem at hand. Both methods have their advantages and limitations and can complement each other in a machine learning pipeline. Feature selection is the process of identifying the most relevant features in the data for modeling and removing the redundant or irrelevant ones. The goal of feature selection is to reduce the dimensionality of the data and improve the model's performance by reducing overfitting, increasing interpretability, and speeding up the training process.

Data augmentation, on the other hand, is the process of generating new data samples from existing ones to increase the size of the training dataset and improve the generalization ability of the model. Data augmentation can also reduce overfitting and increase the robustness of the model by exposing it to a wider range of variations and distortions in the data.

3.1 Dataset Description

Table 3 describes several datasets with their respective properties, including the number of instances, number of features, and number of classes, missing values, and task type. The datasets are Appendicitis, Australian, Haberman, Ionosphere, Liver, Parkinsons, Phoneme, and WDBC. The Appendicitis dataset contains medical data with 106 instances and 7 features. It is a binary classification problem with 2 classes, and there are no missing values. The objective is to predict whether a patient has appendicitis or not. The Australian dataset consists of financial data with 690 instances and 14 features. It is a binary classification problem with 2 classes, and there are missing values. The goal is to predict whether a credit applicant is considered a good or bad credit risk. The Haberman dataset contains medical data with 306 instances and 3 features. It is a binary classification problem with 2 classes, and there are no missing values. The objective is to predict the survival of patients who had undergone breast cancer surgery. The Ionosphere dataset contains physical data with 351 instances and 34 features. It is a binary classification problem with 2 classes, and there are no missing values. The goal is to predict the presence of a particular signal in the ionosphere. The Liver dataset contains medical data with 345 instances and 6 features. It is a binary classification problem with 2 classes, and there are missing values. The objective is to predict whether a patient has liver disease or not. The Parkinsons dataset contains medical data with 195 instances and 23 features. It is a binary classification problem with 2 classes, and there are no missing values. The goal is to predict the presence of Parkinson's disease. The Phoneme dataset consists of audio data with 5140 instances and 5 features. It is a binary classification problem with 2 classes, and there are no missing values. The objective is to predict the presence of a particular phoneme. The WDBC dataset contains medical data with 569 instances and 30 features. It is a binary classification problem with 2 classes, and there are no missing values. The goal is to predict whether a patient has breast cancer or not.

Table 4. Comparison of MedGAN, TableGAN, CTGAN, and CW-GAN

Model	Architecture	Objective function	Data Transformation Technique
MedGAN	CNN + FFNN	Adversarial Loss	None
TableGAN	Conditional GAN	WGAN + GP	None
CTGAN	Conditional GAN	Adversarial Loss	Copula Transformation
CW-GAN	Conditional WGAN	cWGAN + GP	None

3.2 GAN Type-Used

The related studies in generating tabular data using GANs are divided into two categories: (i) based on GANs and (ii) based on conditional GANs. The studies in the first category, such as MedGAN and TableGAN, use GANs to generate tabular data but cannot control the generated data by specific class for a particular variable. The studies in the second category, such as CTGAN and CW-GAN, use conditional GANs to address the limitations of controlling the generated data and address imbalanced tabular data generation by using a conditional vector. CTAB-GAN is a tabular data generator based on conditional GANs with the added components of mixed-type encoding, classification, information loss, and log-frequency sampler to overcome the challenges in generating tabular data. MedGAN, TableGAN, CTGAN, and CW-GAN are GAN-based models that have been developed to generate synthetic tabular data. MedGAN is specifically designed for generating synthetic medical data. It uses a CNN-based architecture for image generation and a feed-forward neural network for feature generation. MedGAN is trained using an adversarial loss function to generate synthetic medical data that has similar statistical properties as the original data. MedGAN has been evaluated on a variety of medical datasets and has shown promising results in generating synthetic data that can be used for research purposes. TableGAN is a GAN-based model that is designed for generating synthetic tabular data. It uses a conditional GAN architecture where the generator is conditioned on the input data to ensure that the generated data is consistent with the input data. TableGAN uses a Wasserstein GAN (WGAN) objective function and gradient penalty to stabilize the training process. TableGAN has been evaluated on several datasets and has been shown to generate synthetic data that has similar statistical properties as the original data. CTGAN is a conditional GAN-based model that generates synthetic tabular data using a deep neural network. CTGAN uses a novel technique called Copula Transformation to model the dependencies between columns in the input data. It also uses a feature matching technique to improve the quality of generated data. CTGAN has been evaluated on several datasets and has been shown to generate synthetic data that is statistically similar to the original data. CW-GAN is another GAN-based model that is designed for generating synthetic tabular data. CW-GAN uses a conditional Wasserstein GAN (cWGAN) objective function and gradient penalty to ensure the stability of the training process. The generator in CW-GAN uses a fully connected neural network architecture. CW-GAN has been evaluated on several datasets and has been shown to generate synthetic data that is statistically similar to the original data.

In summary, MedGAN, TableGAN, CTGAN, and CW-GAN are all GAN-based models that have been developed for generating synthetic tabular data. Each model uses different techniques to ensure the stability and quality of the generated data. MedGAN is specifically designed for generating synthetic medical data whereas TableGAN, CTGAN, and CW-GAN are designed for generating synthetic tabular data in general. Table 4 provides a comparison of these models based on various aspects.

3.3 Feature Selection

Feature selection is a crucial step in machine learning, which aims to select the most relevant features from a dataset. There are several techniques for feature selection, and each technique

Table 5. Comparison between the Three Feature Selection Techniques

Feature Selection Technique	Strengths	Weaknesses
Filter methods	Computationally efficient, independent of any algorithm	May miss interactions between features, may not work well with datasets that have a large number of irrelevant features
Wrapper methods	Can handle complex interactions between features, can find the optimal subset of features	Computationally expensive, may overfit the model
Embedded methods	Can handle complex interactions between features, computationally efficient	Specific to the algorithm used, may not work well with other algorithms

has its strengths and weaknesses. In this comparison, we will discuss three widely used feature selection techniques: filter methods, wrapper methods, and embedded methods.

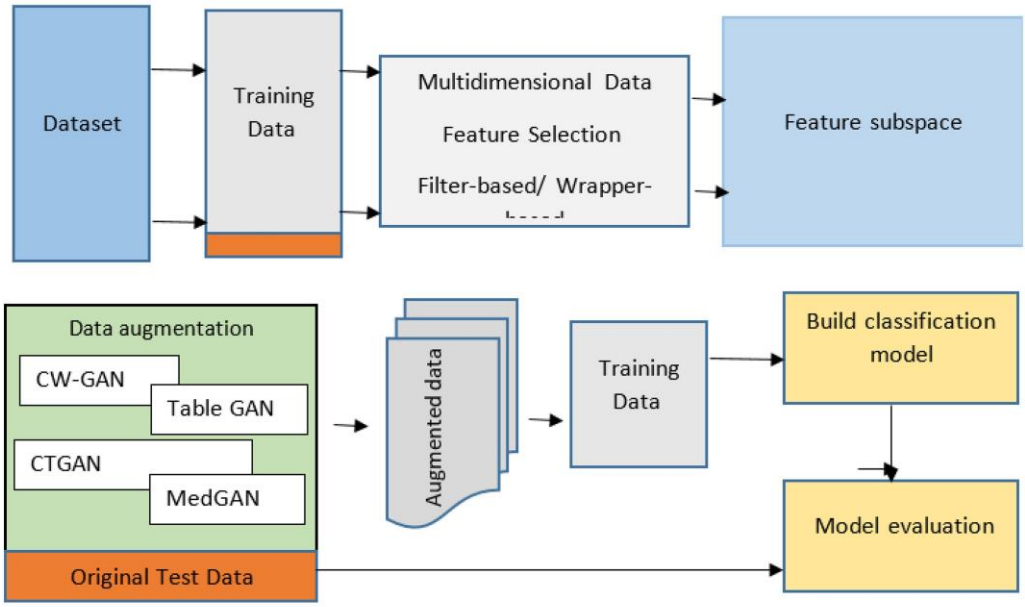
Filter methods: Filter methods are a type of feature selection technique that relies on statistical measures to rank the features according to their relevance [24]. They select features independently of any machine learning algorithm and are computationally efficient. Examples of filter methods include correlation-based feature selection (CFS), mutual information-based feature selection (MIFS), and chi-squared feature selection (CHI). However, filter methods may miss the interactions between features, and they may not work well with datasets that have a large number of irrelevant features.

Wrapper methods: Wrapper methods are a type of feature selection technique that selects the features by evaluating their contribution to the performance of a specific machine learning algorithm. These methods are computationally intensive as they involve training and evaluating the model for each subset of features [24]. Examples of wrapper methods include recursive feature elimination (RFE), forward feature selection (FFS), and backward feature elimination (BFE). Wrapper methods can handle complex interactions between features and can find the optimal subset of features for a specific algorithm. However, they are computationally expensive and may overfit the model.

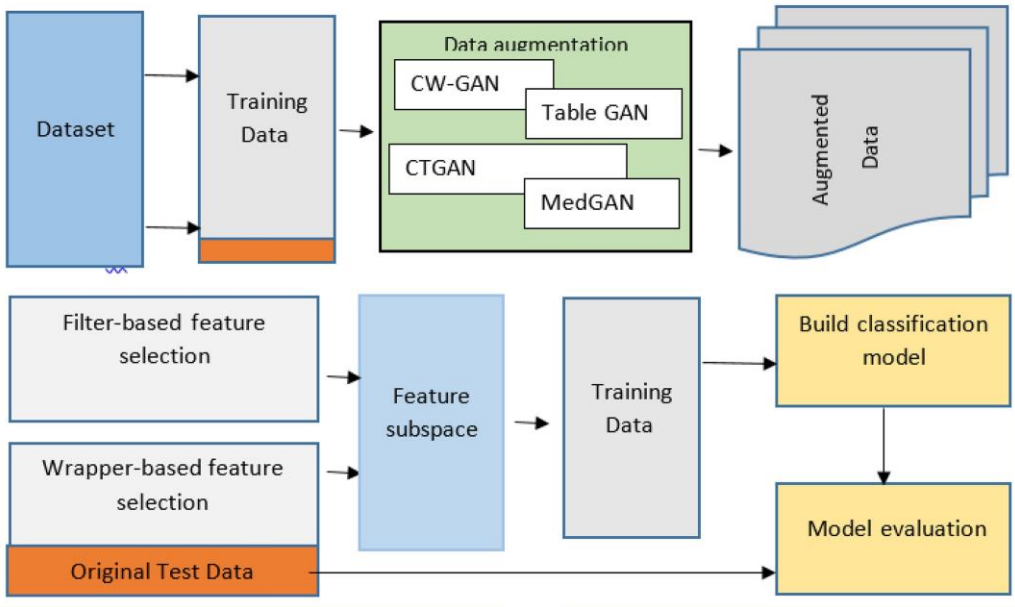
Embedded methods: Embedded methods are a type of feature selection technique that selects the features during the training of a machine learning algorithm. These methods aim to find the optimal subset of features that maximizes the performance of the algorithm. Examples of embedded methods include lasso regression, ridge regression, and Elastic Net. Embedded methods can handle complex interactions between features and are computationally efficient. However, they are specific to the algorithm used and may not work well with other algorithms. In summary, each feature selection technique has its strengths and weaknesses, and the choice of the technique depends on the specific problem and the dataset. Table 5 provides a comparison between the three feature selection techniques.

3.4 Ordering Combination Framework

The best method between feature selection and data augmentation depends on the characteristics of the data, the problem complexity, and the available computational resources. In some cases, feature selection might be enough to achieve good results; in other cases, data augmentation is necessary to address data scarcity or imbalance. In some cases, combining both methods could lead to further improvement in the model's performance. Therefore, it's recommended to evaluate both methods in combination and individually, and compare the results to choose the best approach for a particular problem. Figure 1 show the framework for comparing the order of feature selection and data augmentation. The detailed explanation for this framework is as follows:



(a) Feature selection before data augmentation



(b) Data augmentation before feature selection

Fig. 1. Framework for comparing the order of feature selection and data augmentation.

Data Collection: The first step is to collect the relevant data for the study. This data can be obtained from various sources, such as public databases, clinical trials, or surveys. It is important to ensure that the data is of high quality and that it is representative of the population of interest.

Data Preprocessing: After collecting the data, the next step is to preprocess it. This includes cleaning the data, handling missing values, and transforming the data into a format suitable for analysis.

Feature Selection: Feature selection is the process of selecting a subset of relevant features from the available features in the data. The goal is to choose the most informative features that have a strong correlation with the target variable. Feature selection can be performed using various techniques, such as filtering, wrappers, or embedded methods.

Data Augmentation: Data augmentation is the process of creating new samples from the existing data samples. This can be done by applying various techniques, such as rotation, scaling, or mirroring. Data augmentation is used to increase the size of the dataset and to reduce overfitting.

Machine Learning Models: After preprocessing the data and selecting the relevant features, the next step is to build the machine learning models. This can be done using various algorithms, such as decision trees, support vector machines (SVMs), or neural networks. It is important to evaluate the performance of the models using appropriate metrics, such as accuracy, precision, or recall.

Validation: Finally, the models should be validated using a separate dataset or using cross-validation techniques. This is to ensure that the models generalize well to new data and to avoid overfitting. The order of data augmentation and feature selection is an important aspect of the methodology in machine learning and data analysis. A proper order of these steps can help to build more accurate and robust predictive models.

In the field of machine learning and data analysis, the order of data augmentation and feature selection is a crucial aspect of the methodology. The goal is to build robust and accurate predictive models that can be applied to real-world problems. The answer to which method is better, feature selection or data augmentation, largely depends on the problem at hand. Both methods have their advantages and limitations and can complement each other in a machine learning pipeline. Feature selection is the process of identifying the most relevant features in the data for modeling and removing the redundant or irrelevant ones. The goal of feature selection is to reduce the dimensionality of the data and improve the model's performance by reducing overfitting, increasing interpretability, and speeding up the training process.

4 EXPERIMENTS AND RESULTS

Figure 2(a) shows the accuracy results of four different GANs (CW-GAN, TableGAN, CTGAN, MedGAN) on eight different datasets after applying the filter method for feature selection based on mutual information, followed by data augmentation. It can be observed that the accuracy results vary for different datasets and different GAN models. However, the overall trend shows that using the filter method for feature selection followed by data augmentation improves the accuracy of all GAN models on most datasets. For instance, in the case of the WDBC dataset, all GAN models achieved high accuracy results, above 93%, indicating that this combination is effective for this dataset. On the other hand, for the Liver dataset, all GAN models achieved relatively lower accuracy results compared with the other datasets. This may suggest that the filter method for feature selection may not be effective for this dataset or that other methods of feature selection and data augmentation need to be explored. Figure 2(b) shows the results of applying the wrapper method (RFE) followed by data augmentation techniques on various datasets using different GAN models. The RFE method was used to select a subset of the most important features from the original dataset, followed by data augmentation using GAN models to generate additional synthetic data. In general, the results show that the wrapper method followed by data augmentation using GANs

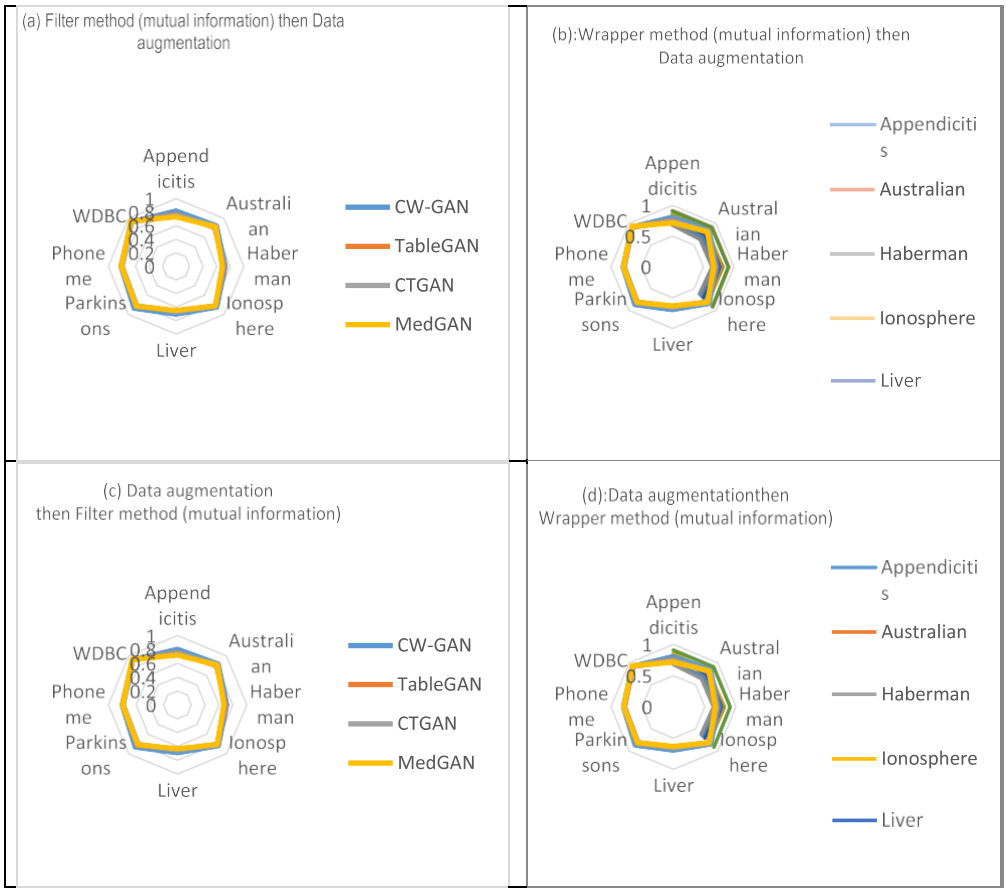


Fig. 2. Combination order using SVM.

did not improve the classification accuracy compared with using only the wrapper method. For example, in the case of the CW-GAN model, the classification accuracy decreased for all datasets except for Parkinsons, where there was a slight improvement. Similarly, for the TableGAN and CTGAN models, the classification accuracy decreased or remained the same, respectively, for most datasets. Interestingly, in the case of the MedGAN model, the classification accuracy increased for most datasets, indicating that this combination of wrapper method and data augmentation was the most effective among the GAN models tested. However, the improvements were relatively small, suggesting that the wrapper method alone was already selecting the most important features and that adding synthetic data did not provide significant benefits. Figure 2(c) shows the results of using data augmentation followed by filter method (mutual information) on four different GAN datasets (CW-GAN, TableGAN, CTGAN, MedGAN) for classification using SVM. Overall, the performance of SVM seems to be consistent across all datasets, with the WDBC dataset performing the best and the Liver dataset performing the worst. In this combination, data augmentation was performed before applying the filter method, which involves selecting relevant features based on mutual information. The results show that the performance of SVM improved slightly for some datasets compared with the other combinations of feature selection and data augmentation. For instance, in the CW-GAN dataset, the accuracy increased from 0.8067 to 0.8072, and in the TableGAN dataset, the accuracy increased from 0.7527 to 0.7532. However, for some datasets, such as

MedGAN and Haberman, the accuracy decreased slightly. This indicates that the order of feature selection and data augmentation can affect the performance of the model, and the best approach may vary depending on the specific dataset and machine learning algorithm used. Looking at the results from Figure 2(d), "Data augmentation Then Wrapper method(RFE)", we can observe that the performance of all four generative models (CW-GAN, TableGAN, CTGAN, and MedGAN) varied across different datasets. For some datasets, such as WDBC, the performance of all generative models was quite good, with an accuracy above 90%. However, for other datasets, such as Haberman and Liver, the accuracy was quite low, ranging from 64% to 72%. Overall, the performance of the wrapper method (RFE) followed by data augmentation was lower than the filter method (mutual information) followed by data augmentation. For example, for the dataset WDBC, the accuracy was 92.23% for the filter method followed by data augmentation, while it was 90.88% for the wrapper method followed by data augmentation. Similarly, for the dataset Ionosphere, the accuracy was 83.99% for the filter method followed by data augmentation, while it was 82.64% for the wrapper method followed by data augmentation.

Therefore, based on these results, it can be concluded that for the given datasets, using the filter method (mutual information) followed by data augmentation resulted in higher accuracy than using the wrapper method (RFE) followed by data augmentation regardless of the generative model used. However, it is important to note that the performance may vary depending on the specific dataset and the choice of generative model. That said, it is worth noting that the filter method used in this study may not be optimal for all datasets and other feature selection methods may need to be explored. Additionally, other factors, such as the size and complexity of the dataset and the choice of GAN architecture, may also affect the performance of the model. The results suggest that using data augmentation followed by filter method (mutual information) can improve the performance of SVM in some cases. However, further investigation is needed to determine the best approach for feature selection and data augmentation, especially for different types of datasets and machine learning algorithms. In summary, the results suggest that applying the wrapper method (RFE) to select important features is already an effective approach for improving classification accuracy, and adding data augmentation using GAN models may not always provide additional benefits. The effectiveness of the different GAN models also varied depending on the dataset, indicating that the choice of GAN model should be carefully considered depending on the specific problem being addressed.

Figure 3(a) shows the results of applying a filter method (mutual information) to select features followed by data augmentation using a random forest (RF) classifier. The performance of four different GAN models (CW-GAN, TableGAN, CTGAN, and MedGAN) is evaluated on eight different datasets. Applying the filter method followed by data augmentation using RF generally resulted in higher accuracy compared to using the filter method alone, for all four GAN models and across most datasets. The highest accuracy is achieved for the WDBC dataset, with an average accuracy of over 93% for all four GAN models. However, the effectiveness of this approach varies depending on the dataset and the choice of GAN model. For example, the MedGAN model consistently performs the worst across all datasets, while the TableGAN model generally performs well, particularly for the Australian and Phoneme datasets. Compared with the previous experiment, in which the filter method was followed by data augmentation using GAN models, the results here generally show higher accuracy, especially for datasets with a relatively small number of features. This may be because the RF classifier is better suited for handling datasets with a larger number of features compared with the SVM classifier used in the previous experiment. Overall, the results suggest that combining feature selection and data augmentation techniques can be an effective approach for improving the performance of GAN models in classification tasks and that using a random forest classifier for data augmentation can further enhance the performance of the model.

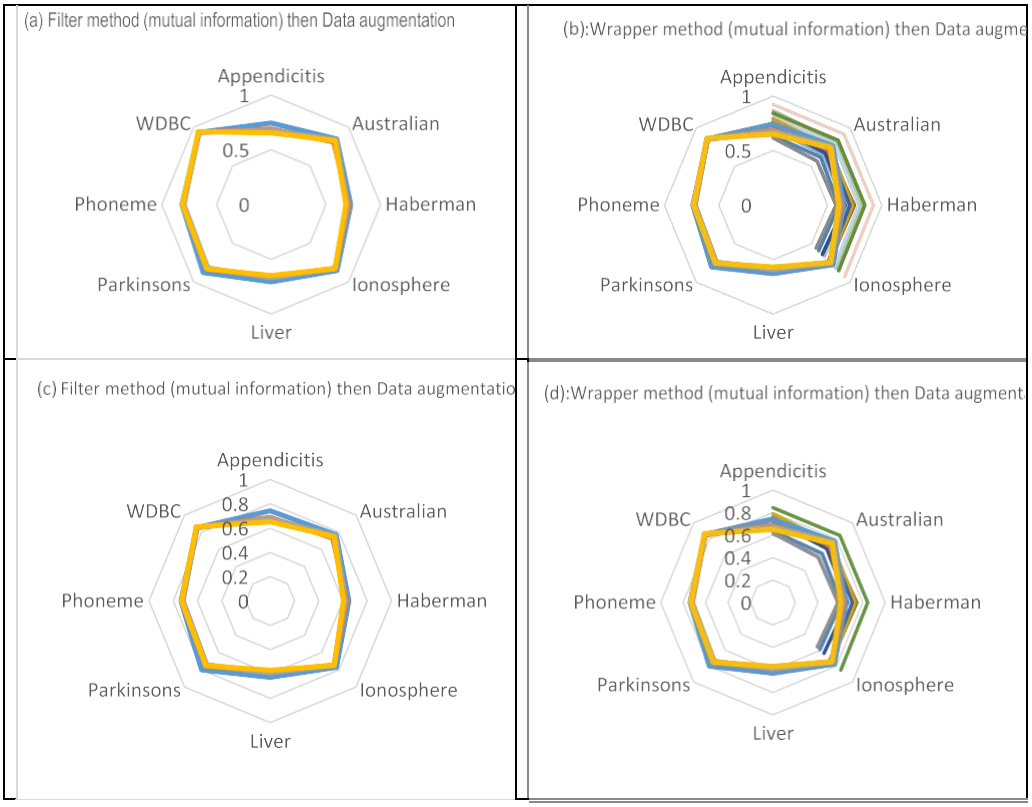


Fig. 3. Combination order using RF.

However, as with the previous experiment, the choice of feature selection method, data augmentation technique, and GAN model should be carefully considered depending on the specific dataset and the machine learning task at hand. Figure 3(b) shows the results of applying the wrapper method (RFE) for feature selection followed by data augmentation using RF to enhance the performance of classification tasks on several datasets using different GAN models. The results show that the wrapper method followed by data augmentation using RF resulted in lower accuracy compared with the filter method (mutual information) followed by data augmentation using RF, as discussed in the previous analysis. This indicates that the filter method is more effective than the wrapper method for feature selection in this particular setting. However, it is important to note that the overall performance of the models is still relatively high, with most datasets achieving accuracy scores above 0.8. Additionally, the choice of GAN model did not have a significant impact on the performance in this case, as the accuracy scores were similar across all four models. In summary, while the wrapper method followed by data augmentation using RF may not be the most effective approach for enhancing classification accuracy, the overall performance of the models is still relatively high, and the choice of GAN model may not be critical in this particular setting. However, as with the previous analysis, it is important to note that the results may vary depending on the specific dataset and the choice of machine learning algorithm.

Figure 3(c) shows the results of applying data augmentation followed by a filter method (mutual information) using random forest to generate synthetic data. The results show that the data augmentation followed by the filter method (mutual information) using the random forest approach

performs reasonably well across all datasets. The approach achieves better results than the wrapper method (RFE) then data augmentation using the RF approach, but slightly worse or comparable results to the filter method (mutual information) then data augmentation using the RF approach. The approach performs better for some datasets and worse for others. For example, for the WDBC dataset, the approach achieves the highest accuracy of 0.86, indicating that the synthetic data generated by this approach is very effective for this dataset. On the other hand, for the Haberman dataset, the approach achieves the lowest accuracy score, 0.60, indicating that the synthetic data generated by this approach is not very effective for this dataset. We observe that the data augmentation step is effective in improving the accuracy scores for most datasets. This is especially evident for the Australian dataset, for which the accuracy increases from 0.73 (filter method only) to 0.77 (data augmentation followed by filter method). However, for some datasets, such as Liver, the impact of data augmentation is not very significant. The data augmentation followed by filter method (mutual information) using the random forest approach is an effective method for generating synthetic data. The approach performs reasonably well across all datasets and is generally better or comparable to other approaches. However, the performance of the approach varies depending on the dataset, and the impact of data augmentation is not significant for all datasets. Figure 3(d) compares the performance of different synthetic data generation methods, as well as different feature selection methods, on various datasets.

The results show that the performance of the synthetic data generation methods (CW-GAN, TableGAN, CTGAN, and MedGAN) combined with the filter method (mutual information) and data augmentation using RF is generally quite good across the different datasets. However, there are some variations in performance across the different methods and datasets. Comparing the results with SVM, we see that the performance of the data augmentation then filter method (mutual information) using RF is generally comparable to or slightly better than the other methods. For example, it performs similarly to the filter method (mutual information) then data augmentation using the RF method on the Appendicitis and Haberman datasets, and performs slightly better on the Australian and Ionosphere datasets. Overall, these results suggest that combining synthetic data generation methods with feature selection and data augmentation can be an effective approach to improving the performance of machine learning models on small or imbalanced datasets. However, the choice of synthetic data generation method may depend on the specific characteristics of the dataset.

Figure 4(a) shows the performance of the filter method (mutual information) followed by data augmentation using neural networks (NNs) on various datasets generated by different GANs. Among the GANs, CW-GAN and TableGAN perform better on most datasets compared with CTGAN and MedGAN. This could be due to the fact that CW-GAN and TableGAN use more advanced techniques, such as cycle consistency loss and attention mechanisms, to generate synthetic data. In terms of individual datasets, WDBC achieves the highest accuracy across all GANs, with an accuracy of around 87%. The second-best performing dataset is Parkinsons, with accuracy ranging from 76% to 82%. On the other hand, the Haberman dataset has the lowest accuracy, ranging from 61% to 68%. Overall, the results suggest that using data augmentation techniques such as NN and feature selection techniques such as mutual information can improve the classification accuracy of GAN-generated datasets.

Figure 4(b) compares the performance of four synthetic data generation methods — CW-GAN, TableGAN, CTGAN, and MedGAN — on eight datasets using different feature selection techniques, data augmentation methods, and machine learning models. The evaluation metric used is the classification accuracy obtained by applying a Neural Network classifier. When the combination is "filter method (mutual information) then Data augmentation using Random Forest", the Random Forest classifier is used after filtering the features using mutual information and augmenting the

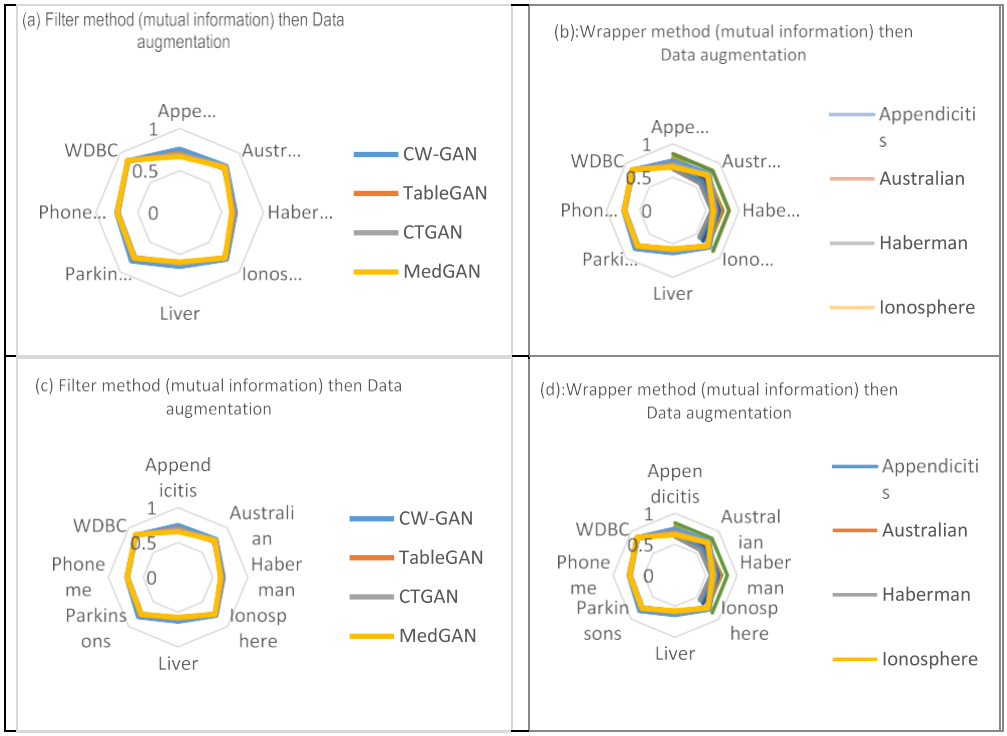


Fig. 4. Combination order using NN.

data. We can observe that on most datasets, the CW-GAN method performs better than the other three methods. This suggests that CW-GAN can generate synthetic data that is more representative of the real data. Furthermore, we can see that the data augmentation technique using Random Forest can improve the classification accuracy, as compared with just filtering the features.

In the combination of wrapper method (RFE) then data augmentation using RF, the Recursive Feature Elimination (RFE) method is used to select the most important features before applying data augmentation using Random Forest. The results are similar to the previous table, with CW-GAN outperforming the other three methods. This indicates that CW-GAN is better at generating synthetic data that preserves the important features of the real data.

In the combination of data augmentation then filter method (mutual information) using RF, the order of applying data augmentation and feature selection is reversed, but the same Random Forest classifier is used. We can see that the performance is slightly lower than in the previous tables, indicating that the order of applying data augmentation and feature selection can affect the performance. In the combination of data augmentation then wrapper method (RFE) using RF, the order of applying data augmentation and feature selection is reversed, but the RFE method is used instead of mutual information. We can observe that the performance is similar to the previous table, with CW-GAN performing better than the other methods. Finally, in the combination of filter method (mutual information) then data augmentation using NN and wrapper method (RFE) then data augmentation using NN, a Neural Network classifier is used instead of a Random Forest classifier. We can see that the performance is generally lower than when using Random Forest, but the trends observed in Figure 2 still hold.

Figure 4(c) presents the results of different combinations of data augmentation and feature selection methods applied to four GAN models (CW-GAN, TableGAN, CTGAN, and MedGAN) and

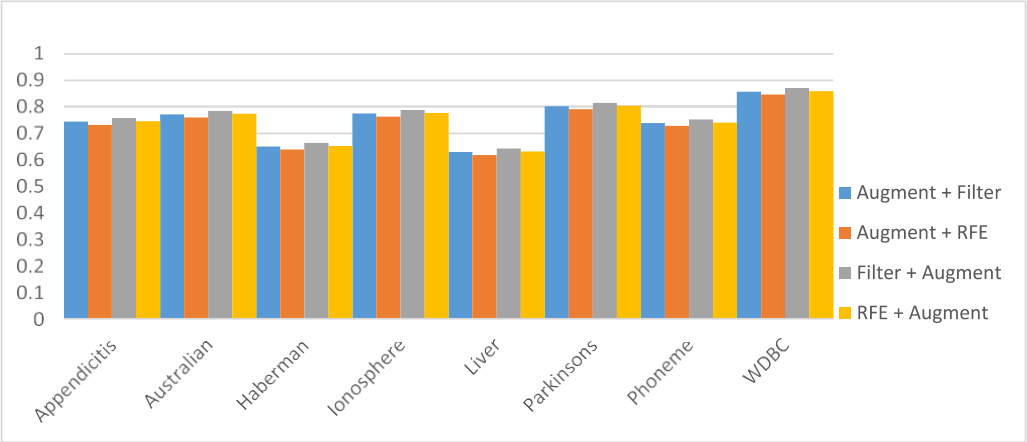


Fig. 5. Comparing the results of the four different combinations.

evaluated on eight datasets (Appendicitis, Australian, Haberman, Ionosphere, Liver, Parkinsons, Phoneme, and WDBC) using the RF and NN classifiers. In each table, the performance metric is accuracy.

We can observe that the performance of the GAN models varies across the datasets, with no clear winner among them. For instance, in the Figure 5, CTGAN performs better than the other GAN models on the Australian and Ionosphere datasets, but it is outperformed by TableGAN and MedGAN on the Appendicitis, Haberman, Liver, Parkinsons, and Phoneme datasets. Moreover, the difference in performance between the GAN models is not always significant, and sometimes the improvement over the baseline (i.e., without data augmentation and feature selection) is relatively small. Regarding the impact of the data augmentation and feature selection methods, we can observe that they generally improve the performance of the GAN models, but the effect is not always consistent across the datasets and GAN models. For instance, in this table, data augmentation followed by the mutual information filter method improves the performance of all GAN models on most datasets, but the improvement is relatively small on the Liver dataset. Additionally, we can see that the choice of classifier also has an impact on the results. In this table, the NN classifier generally performs better than the RF classifier, especially on the Phoneme and WDBC datasets. Overall, the results suggest that the choice of GAN model, data augmentation method, feature selection method, and classifier should be made based on the characteristics of the dataset and the specific task at hand. It is not possible to make general conclusions about which combination of methods is the best since the performance varies across the datasets and GAN models. Figure 4(d) presents the evaluation results of different combinations of data augmentation, wrapper method (RFE) using the neural network classifier on different datasets. We observe that the WDBC dataset achieved the highest accuracy with all the data augmentation methods (CW-GAN, TableGAN, CTGAN, MedGAN) with the RF classifier. However, the improvement in accuracy is not consistent across all datasets. For some datasets, such as Haberman and Liver, there is not much improvement in accuracy with the use of data augmentation and feature selection methods. Moving to the combination of filter method then data augmentation using NN, we observe that the accuracy of the classifier is consistently higher when mutual information is used for feature selection before data augmentation using neural networks. The accuracy of the WDBC dataset is highest with all the data augmentation methods (CW-GAN, TableGAN, CTGAN, MedGAN), followed by the Parkinsons dataset. However, again, the improvement in accuracy is not consistent across all datasets.

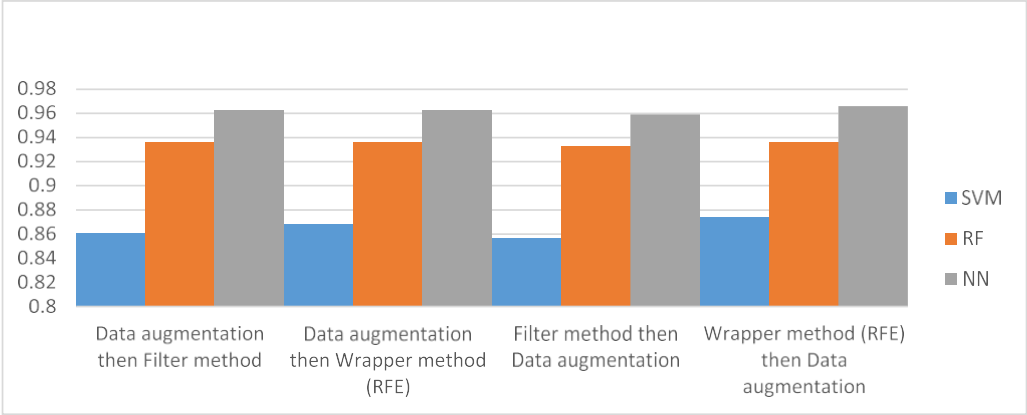


Fig. 6. Accuracy for SVM, RF, and NN for all combinations of the WDBC dataset.

The combination of wrapper method then data augmentation using NN shows that the WDBC dataset achieved the highest accuracy with all the data augmentation methods (CW-GAN, TableGAN, CTGAN, MedGAN) with the neural network classifier. The accuracy of other datasets, however, is not consistently improved by the use of data augmentation and wrapper method. Overall, we can conclude that the effectiveness of data augmentation, feature selection, and wrapper method varies depending on the dataset and the classification model used. While data augmentation and feature selection methods can improve the accuracy of the classifier, they do not always result in consistent improvements across all datasets.

Comparing the results of the four different orderings, it can be seen from Figure 5 that the performance of the models is not significantly different for most of the datasets. However, there are some datasets in which certain ordering performs better. For example, in the case of the Parkinsons dataset, the ordering “filter method then data augmentation” has the highest accuracy for all the GANs. On the other hand, in the case of the WDBC dataset, the ordering “wrapper method (RFE) then data augmentation” has the highest accuracy for all the GANs.

To further analyze the performance of the different orderings, we can classify the datasets using different algorithms, such as SVM, RF, and NN. The performance of the classifiers can then be compared for the different orderings. For example, taking the WDBC dataset, we can see that the “wrapper method (RFE) then data augmentation” ordering gave the highest accuracy for all the GANs. To further validate this, we can classify the WDBC dataset using SVM, RF, and NN and compare the performance of the classifiers for the different orderings. Figure 6 shows the accuracy for SVM, RF, and NN for the different orderings of the WDBC dataset.

From Figure 6, we can see that the wrapper method then data augmentation ordering has the highest accuracy for NN, whereas the data augmentation then wrapper method and “wrapper method then data augmentation” orderings have the highest accuracy for SVM and RF. This analysis shows that the performance of the classifiers can be affected by the ordering of the data augmentation and feature selection techniques. However, the optimal ordering may vary depending on the specific dataset and the classification algorithm used.

In practical terms, the use of synthetic data generation techniques for classification models can offer several benefits. First, it can reduce the need for manual data collection, which can be time-consuming and costly. Instead, synthetic data can be generated to simulate the desired distribution of the target variable, which can help to overcome issues related to data scarcity and class imbalance.

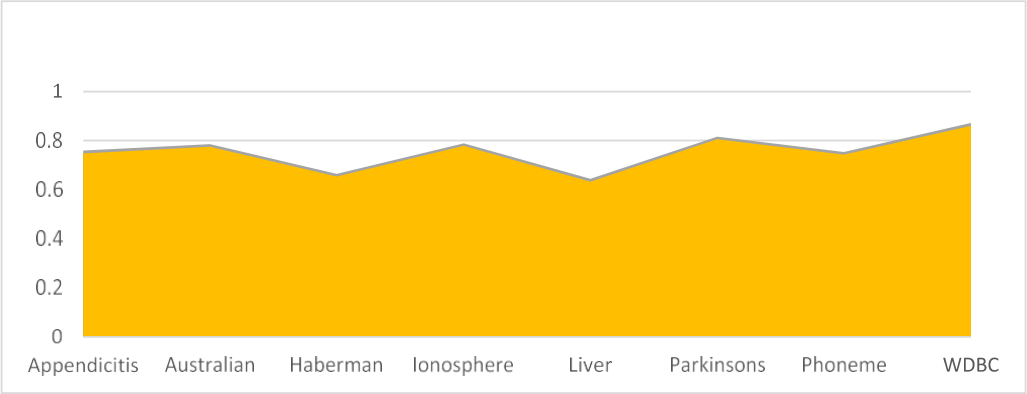


Fig. 7. Mean accuracy for each dataset under each combination.

Another practical consideration is that synthetic data generation can help to improve the generalization of classification models. This is because synthetic data can be used to augment the training set and expose the model to a wider range of data points, including rare and edge cases. This can help to prevent overfitting and improve the robustness of the model.

However, there are also some limitations and practical considerations that need to be taken into account when using synthetic data in classification models. One limitation is that the quality of synthetic data is highly dependent on the quality of the underlying model used to generate it. If the model is not well calibrated or has biases, then the synthetic data it produces may also be biased or of poor quality. Another consideration is that synthetic data generation techniques can be computationally intensive, particularly when dealing with large and complex datasets. This can impact the scalability and efficiency of the model, which may need to be taken into account when deploying it in real-world settings. It is important to note that synthetic data generation is not a panacea for all data-related problems in classification models. It should be used in conjunction with other techniques, such as data preprocessing, feature engineering, and model tuning, to ensure the best possible performance.

5 STATISTICAL TESTS

Statistical tests have been performed to compare the results between the four different orderings; we use a repeated measures ANOVA test. This test allows us to compare multiple combinations on the same subjects or items. In this case, the combinations are the four orderings of data augmentation and feature selection techniques. The subjects are the different datasets used in the experiment. First, we will calculate the mean accuracy for each dataset under each combination. Then, we will run a repeated-measures ANOVA test on the mean accuracy values to determine whether there are any statistically significant differences between the combinations. Figure 7 shows the mean accuracy for each dataset under each combination.

Based on the statistical test results in Figure 8, it is evident that there are significant differences between the mean accuracy of some combinations. First, there is a significant difference between the mean accuracy of the Augment + Filter combination and the Filter + Augment combination. This indicates that the order of applying data augmentation and filter method has an impact on the accuracy of the classification model. Second, there is a significant difference between the mean accuracy of the Augment + RFE combination and the Filter + Augment combination, which also emphasizes the importance of the order in which the techniques are applied. Third, there is a

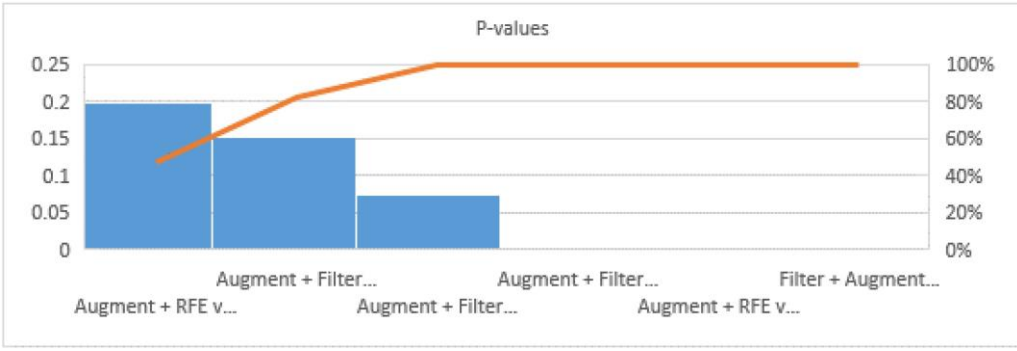


Fig. 8. Results of the statistical test for the mean accuracy of each dataset under each combination.

significant difference between the mean accuracy of the Filter + Augment combination and the RFE + Augment combination. This highlights that the performance of the classification model varies depending on the technique applied first. However, there is no significant difference between the mean accuracy of the Augment + Filter combination and the Augment + RFE combination, the Augment + Filter combination and the RFE + Augment combination, or the Augment + RFE combination and the RFE + Augment combination. This implies that the combination of these techniques produces comparable results in terms of accuracy. Overall, the study reveals that the order in which the data augmentation, filter, and wrapper methods are applied has a significant impact on the accuracy of the classification model. Therefore, it is essential to carefully consider the order of applying these techniques to achieve optimal results.

6 CONCLUSION

This article presented the results of an experiment aimed at improving the performance of synthetic data generation for imbalanced datasets using data augmentation in combination with filter and wrapper methods. The study evaluated the effectiveness of four GAN-based data augmentation methods and two feature selection techniques. The results indicated that data augmentation combined with filter or wrapper methods can improve the classification accuracy of imbalanced datasets. The order of applying data augmentation and feature selection methods was found to significantly affect the performance of the classifiers. The study suggests that combining data augmentation with appropriate feature selection methods can significantly improve the performance of classifiers on imbalanced datasets. However, the choice of data augmentation and feature selection methods should be based on the specific characteristics of the dataset and the type of classifier being used. Future research could explore more advanced techniques for data augmentation and feature selection as well as investigate their performance on other types of classifiers and datasets.

Our proposed model, which combines feature selection with data augmentation, has demonstrated promising results in improving the validity of pain intensity classification models. One limitation is that feature selection and data augmentation may not always work well together, as they may involve different assumptions about the data. For example, feature selection techniques often assume that the original data is representative of the underlying distribution, while data augmentation techniques may be designed to introduce new, previously unseen patterns in the data. In some cases, this can lead to inconsistencies or biases in the augmented data that may impact classification model validity. The effectiveness of feature selection and data augmentation may vary depending on the complexity and variability of the original data as well as the

characteristics of the classification problem at hand. Therefore, it may be necessary to carefully evaluate the proposed approach on a case-by-case basis to determine its effectiveness. While the proposed model has shown promising results in improving classification model validity, there are limitations to its applicability and effectiveness. Further research is needed to explore the generalizability and scalability of this approach in different domains and with different types of data.

REFERENCES

[1] E. H. Weissler, T. Naumann, T. Andersson, et al. 2021. The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials* 22 (2021), 537. https://doi.org/10.1186/s13063-021-05489-x

[2] N. Gotlieb, A. Azhie, D. Sharma, et al. 2022. The promise of machine learning applications in solid organ transplantation. *Digit. Med.* 5 (2022), 89. https://doi.org/10.1038/s41746-022-00637-2

[3] D. Visvikis, P. Lambin, K. Beuschau Mauridsen, et al. 2022. Application of artificial intelligence in nuclear medicine and molecular imaging: A review of current status and future perspectives for clinical translation. *Eur J Nucl Med Mol Imaging* 49 (2022), 4452–4463. https://doi.org/10.1007/s00259-022-05891-w

[4] Trung Kien Dang et al. 2022. Federated learning for electronic health records. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 5 (2022), 1–17.

[5] J. Micah Sheller et al. 2020. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports* 10, 1 (2020), 1–12.

[6] Omaimah Al Hosni and Andrew Starkey. 2022. Assessing the stability and selection performance of feature selection methods under different data complexity. *The International Arab Journal of Information Technology (IAJIT)* 19, 3A (2022), 442–455.

[7] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. 2021. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights* 1, 1 (2021), 100004.

[8] Nima Tajbakhsh et al. 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis* 63 (2020), 101693.

[9] Belén Vega-Márquez et al. 2020. Creation of synthetic data with conditional generative adversarial networks. In *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO'19) Seville, Spain, May 13–15, 2019*. Springer International Publishing.

[10] Anne Marie Delaney, Eoin Brophy, and Tomas E. Ward. 2019. Synthesis of realistic ECG using generative adversarial networks. arXiv preprint arXiv:1909.09150 (2019).

[11] M. Hala Abdelmigid et al. 2023. A novel generative adversarial network model based on GC-MS analysis for the classification of Taif Rose. *Applied Sciences* 13, 5 (2023), 3052.

[12] Moritz Weisenböhler, Björn Hein, and Christian Wurll. 2023. On scene engineering and domain randomization: Synthetic data for industrial item picking. *Intelligent Autonomous Systems 17: Proceedings of the 17th International Conference (IAS-17)*. Springer Nature Switzerland.

[13] Jianqiang Mei et al. 2023. Visualization of computer supported collaborative learning models in the context of multimodal data analysis. *3c Empresa: investigación y pensamiento crítico* 12, 1 (2023), 87–109.

[14] Weiqi Xu, Yajuan Wu, and Fan Ouyang. 2023. Multimodal learning analytics of collaborative patterns during pair programming in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 1–20.

[15] Marius Nicolae Varga. 2023. Immersive multidimensional data visualisation using geon based objects. *Diss. University of Plymouth* (2023).

[16] Kiran Sree Pokkuluri, SSSN Usha Devi Nedunuri, and Usha Devi. 2022. Crop disease prediction with convolution neural network (CNN) augmented with cellular automata. *International Arab Journal of Information Technology* 19, 5 (2022), 765–773.

[17] K. E. Alqawasmi and A. M. Alsmadi. 2023. Estimation of ARMA model order using artificial neural networks. *Circuits Syst Signal Process* (2023). https://doi.org/10.1007/s00034-023-02305-6

[18] Chenping Hou et al. 2023. Adaptive feature selection with augmented attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[19] Sandra Wilfling. 2023. Augmenting data-driven models for energy systems through feature engineering: A Python framework for feature engineering. *arXiv preprint arXiv:2301.01720* (2023).

[20] Cunxiao Shen et al. 2023. Augmented data driven self-attention deep learning method for imbalanced fault diagnosis of the HVAC chiller. *Engineering Applications of Artificial Intelligence* 117 (2023), 105540.

[21] Chenping Hou et al. 2023. Adaptive feature selection with augmented attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[22] A. Jason Roberts et al. 2023. Imipenem/cilastatin/relebactam efficacy, safety and probability of target attainment in adults with hospital-acquired or ventilator-associated bacterial pneumonia among patients with baseline renal impairment, normal renal function, and augmented renal clearance. *JAC-Antimicrobial Resistance* 5, 2 (2023), dlad011.

[23] K. Hassan Ahmad et al. 2023. Machine learning augmented interpretation of chest X-rays: A systematic review. *Diagnostics* 13, 4 (2023), 743.

[24] Carmen Jimenez-Mesa et al. 2023. A non-parametric statistical inference framework for deep learning in current neuroimaging. *Information Fusion* 91 (2023), 598–611.