Synthetic Versus Real: A Deep Dive into Data Augmentation Techniques for Image Classification

TOBIAS JOHANNESSON
tjohann
@kth.se

January 13, 2024

Abstract

The unprecedented surge in data generation presents a significant challenge to its effective utilization. This research conducts a deep dive into data augmentation techniques for image classification, focusing on Generative Adversarial Networks (GANs) and Synthetic Minority Over-sampling Technique (SMOTE) for improving training robustness in high-dimensional, imbalanced image datasets. We explore how these techniques affect the performance of deep learning models, particularly VGG, using metrics like AUC score, accuracy, recall, and F1-Score. Furthermore, we conduct an analysis of recent data augmentation methodologies, highlighting their potentials and limitations to handle class imbalances and showcasing the potential for enhanced training robustness via GAN-based augmentation.

Our research underscores the strategic importance of integrating data augmentation techniques, to tackle challenges associated with class imbalance in high-dimensional datasets. The use of these techniques impact model generalization, ease-of-use, sustainability, and overall performance. Our findings suggest that while GANs show potential, their effectiveness hinges on data quality and model simplicity, offering insights for future research on optimizing data augmentation strategies.

Data Augmentation, SMOTE, Generative Adversarial Networks, Deep Learning, Machine Learning, Data Science

Contents

1 Introduction

The unprecedented surge in high-dimensional data gives rise to a set of challenges in the domain of data science. This backdrop highlights the issues of data imbalance and scarcity, which impede the full potential of machine learning models, leading to overfitting and diminished predictive accuracy. In this report, we examine these problems, emphasizing the current limitations and knowledge gaps. We then introduce the potential of Generative Adversarial Networks (GANs) in addressing these challenges, acknowledging the need for improved sample quality. This leads us to our research question: How does combining enhanced GAN methodologies with other data augmentation techniques affect the performance of models on complex datasets? This question underpins our study's aim to explore innovative solutions in data augmentation, seeking to advance the efficiency and reliability of machine learning models in handling complex data scenarios.

1.1 Background

The sheer volume of generated data has reached unprecedented levels, and it continues to increase over the decades. Every day, sensors and cameras contribute quintillions of bytes, culminating in an overwhelming abundance. This exponential surge, as highlighted by Zhai et al. [1], leads to the challenge known as the curse of high dimensionality which comes from not only the increase in sheer volume but also from the complexity of this data. Simultaneously, enterprises striving to compete in dynamic business ecosystems contribute substantially to this data deluge, amplifying the challenges faced in handling such vast datasets [2]. This phenomenon presents unique challenges in data processing and analysis, especially in the extraction of meaningful insights from vast, intricate datasets commonly referred to as big data [2].

In all this generated data, high-dimensional data, notably in image form, amplifies the difficulty of efficiently extracting meaningful information. Donoho et al. [3] liken this endeavor to the proverbial search for a needle in a haystack, emphasizing the arduous nature of the task. Analyzing high-dimensional data necessitates exhaustive evaluations, often scaling to tens of trillions, complicating data analysis and rendering traditional algorithms impractical or inefficient [4]. This challenge exacerbates with the propensity of high-dimensional datasets to overfit, compromising model generalization on new, unseen data due to limited sample sizes [4].

1.2 Problem Area

Despite the advancements in computational power and algorithms, the issue of data imbalance and scarcity remains a significant hurdle in harnessing the full potential of machine learning models, particularly in high-dimensional contexts [5, 6]. This imbalance often leads to models overfitting during training, as they learn to mimic the provided data rather than learn the underlying patterns, diminishing their predictive accuracy on new, unseen data [7, 8]. Traditional data augmentation techniques, while beneficial, have shown limitations when applied to high-dimensional data, and they often fail to adequately address the complexity and the nuanced nature of such data, resulting in sub-optimal model performance. This gap in effective data augmentation methods for imbalanced and high-dimensional datasets is a critical problem area that needs further exploration and innovative solutions [9].

Building on these challenges, recent studies have underscored the potential of Generative Adversarial Networks (GANs) in data augmentation, particularly for high-dimensional datasets [10, 11]. However, these studies also highlight critical areas that warrant further investigation. Firstly, the quality of samples generated by GANs and their effectiveness across diverse datasets require enhancement [10]. This involves not only improving the fidelity of generated data but also ensuring its robustness and utility in varied application contexts. Secondly, the difficulty in training GANs, especially with limited or low-dimensional datasets, poses a significant barrier [11]. The performance degradation under these conditions suggests a need for novel approaches in GAN methodologies, possibly through innovative algorithmic improvements

or the integration of GANs with other data augmentation techniques such as multiple GANs, WGANs, or SMOTE.

These gaps present an opportunity for exploration into how GANs, particularly when used in an ensemble with other data augmentation methods, can be optimized for high-dimensional, imbalanced data environments. While existing data augmentation methods have shown efficacy, they often do not fully address the unique complexities of high-dimensional, imbalanced datasets, one of the gaps our research aims to fill. Then also recognizing the computational intensity of GAN methodologies, this study explores complementary techniques, balancing performance with resource efficiency. Therefore, this study aims to delve into these unexplored areas, seeking to harness the potential and explore limitations of current data augmentation strategies for high-dimensional data.

1.3 Research Question

This study seeks to answer the following question: How does combining enhanced GAN methodologies with other data augmentation techniques affect the performance of models on complex datasets? This study explores the synergy between enhanced GAN methodologies and various data augmentation techniques to understand their collective impact on model performance. Specifically, the focus is on handling challenges posed by high-dimensional and imbalanced image datasets. Rather than the standard GAN, this report will examine the Conditional GAN or CGAN, and from here on GAN and CGAN will be used interchangeably. The motivation behind this research question stems from the identified need for more sophisticated data augmentation techniques capable of handling the intricacies of high-dimensional data. By exploring the potential of GANs in this context, this study aims to contribute to the advancement of data science methodologies, particularly in optimizing data augmentation strategies. Although the outcomes associated with GAN-based augmentation are not always predictable, this research venture contributes valuable insights into their potential and limitations. The ultimate goal is to develop a nuanced understanding of the role of GANs in data augmentation and to propose strategies that can help simplify dealing with high-dimensional and imbalanced datasets.

2 Extended Background

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks is a simple two-model method to generate new data similar to the training data at hand. It consists of one model called the generator which creates the data samples, and the second model is called the discriminator. These two models compete in a zero sum game to try and beat each other in their respective task. As the discriminative model tries to guess whether data came from the provided real dataset or from the generative model's output, the generative model is trying to maximize the probability of the discriminative model guessing incorrectly. This method starts by first training the discriminator to correctly classify correct data, and only once the discriminator is done the generator starts. As the generative model produces new sample, they are sent to the discriminator which aims to correctly call out the synthetic data, and the outcome of the discriminator's prediction is shared to both models where the loser updates their weights and this zero-sum game continues until the generator becomes good enough to continuously trick the discriminator.[12]

Similar to other models, GANs are not perfect and suffer from problems such as overfitting and can end up capturing specific variances depending on the training of the two models [13]. It can also suffer from vanishing gradients during training if the model ends up being too strong or it can fail to converge which can be remedied by adding random noise to the discriminator's input [14]. In the end this data augmentation is ideal for high dimensional datasets and can generate high quality images upon successful training of the generator [15].

The Conditional Generative Adversarial Network (CGAN) enhances the standard GAN by introducing an

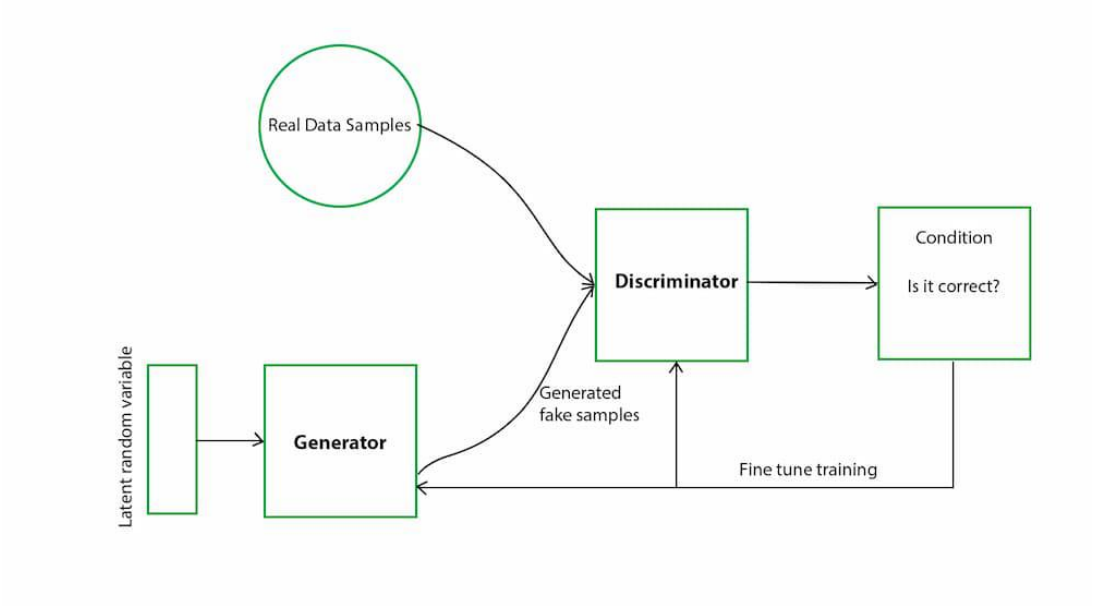


Figure 1: Generative Adversarial Network Diagram from https://www.geeksforgeeks.org/generative-adversarial-network-gan/

additional layer of complexity. In CGAN, extra information is supplied to guide the data generation process. This additional information enables the generation of specific versions of data, such as samples belonging to a particular class. As demonstrated in [16], CGANs offer versatility, allowing for different variations, including subtype CGANs that cater to diverse conditional generation requirements.

2.2 Wasserstein Generative Adversarial Networks (WGANs)

WGANs aim to tackle certain issues found in using GANs such as problems with vanishing gradients and in scenarios where the distribution of generated data and real data differ. WGANs utilized a different loss function called Wasserstein (also known as Earth's mover distance), compared to the GANs Jensen-Shannon divergence, and the discriminator is replaced by a critic. The critic does not identify an image as real or synthetic, but outputs a continuous value on how close it seems to be to data from the real dataset [17]. All this leads to a better range of samples where the model avoids the problem known as model collapse, and by limiting the weights of the critic it avoids the problems of vanishing gradients.

2.3 Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE), an oversampling algorithm designed to address classification challenges in unbalanced datasets [18]. SMOTE artificially creates "synthetic" samples for minority classes, significantly improving accuracy, sensitivity, specificity, and F1 score [19]. The original SMOTE generates new samples randomly on the borders of K-Nearest Neighbours (K-NN) [18]. By identifying the K nearest samples to a randomly selected data point, new data points are created along the border between the data point and its neighbors. Other variations of SMOTE are also proposed by many [8] to improve test performance.

2.4 Ensemble Methods

Ensembles build on the idea that seeking opinions from multiple "experts" usually provides a better decision, and they are successfully adopted in many use cases including unbalanced data [20]. Ensemble methods combine predictions from all trained models to enhance performance and generalisability [21, 22], and [23] presents the idea of using GANs specifically in an ensemble for synthetic training data generation. However, Wen et al. found that combining certain traditional data augmentation techniques with ensemble

techniques can be detrimental to the model performance, in opposition to the prevailing beliefs in current literature [22].

3 Methodology

Our methodology, centered around using a pre-trained VGG model and evaluating with the AUC score, is meticulously chosen to address the core of our research question: assessing the impact of data augmentation techniques on the predictive performance of models on complex data. The VGG model, known for its proficiency in handling complex image data, becomes an ideal candidate for observing the effects of augmentation. The AUC score, a robust metric for performance evaluation, particularly in imbalanced scenarios, offers a clear measure of the efficacy of various augmentation techniques. This combination ensures a focused and effective analysis of how data augmentation can enhance model robustness and accuracy in challenging data environments.

3.1 Choice of Method

Previous research on data augmentation show that the evaluation of generative methods is known to be complex and certain heuristics can provide misleading results [24]. The combination of real and synthetic data used to train machine learning models output AUC scores that could provide a concrete evaluation of different data augmentation techniques and settings where the focus lies on the downstream task for which the synthetic data is generated [11]. In [25] a convolutional neural network used for multi-class classification is trained on the synthetic and real data. Since the original data is unbalanced the accuracy is proven not to be a solid metric for evaluating the performance, and the generated data is instead evaluated using a combination of metrics: Accuracy, Precision, Recall, and F1 Score on the downstream task. This research approach acknowledges the inherent risk that the GAN may not produce synthetic data closely resembling real data, which could potentially skew the conclusions about its impact on the predictive performance of the downstream task. To mitigate this risk, we have adopted a rigorous evaluation framework that includes comparing the results against baseline models trained on real data. Additionally, by transparently publishing our code and methodology, we provide opportunities for future research to validate, critique, or build upon our findings. This open approach not only enhances the credibility of our study but also contributes to the collective advancement of knowledge in the field.

The methodologies presented in the papers by Luzi et al. [26], Zhang et al. [11], and Cui et al. [25] offer valuable insights into the utilization of GANs for data augmentation, each addressing unique challenges pertinent to this field. However, they also present certain limitations, such as potential biases towards less diverse samples known as model collapse, challenges in assessing utility across applications, and scalability issues with complex architectures for high-dimensional data. As stated in [24], the evaluation needs be application specific, and since this report does not focus on the plausibility of synthetic images, or their statistical likeness. This report will use this fact and not examine the image data generated itself, and will only examine how the downstream task is affected.

To address these challenges and effectively answer our research question, our approach adopts the nuanced evaluation approach of Luzi et al. to ensure a comprehensive understanding of the model's performance. Zhang et al.'s focus on generating high-fidelity data closely resembling real datasets will be incorporated into our GAN training process. However, unlike their approach, we will not assess the synthetic data directly, but rather its efficacy in improving model performance, using AUC score and other relevant metrics as robust measures of effectiveness across applications as done in the other two papers. Cui et al.'s work tackles the challenge of balancing datasets, which is crucial for our study and this research draw inspiration from their evaluation techniques. However, to overcome the scalability issues they encountered with complex architectures, we will use a well-known, efficient deep learning model, the VGG network with pre-trained weights to minimze the training required. The use of pre-trained architecture ensures that our approach remains scalable and applicable to high-dimensional data.

3.2 Ethics and Sustainability

In this project, we place a strong emphasis on both ethical and sustainability considerations to ensure responsible research practices. From an environmental standpoint, we are conscious of the energy consumption and climate impact associated with extensive computational processes. To mitigate this, we employ sample data during the initial phases of learning and setting up our experimental suite. Additionally, the use of fine-tuning techniques in the final model training is planned to reduce the computational load and accelerate the training process, further minimizing energy usage.

Ethically, we are committed to using data responsibly. The datasets employed in this research are sourced from publicly available repositories, ensuring that they are collected and shared in compliance with relevant data protection and privacy regulations. This approach not only upholds the ethical standards of data usage but also mitigates risks related to personal or sensitive information. Furthermore, we ensure that the data used does not perpetuate biases or unfairness, maintaining the integrity and societal relevance of our research.

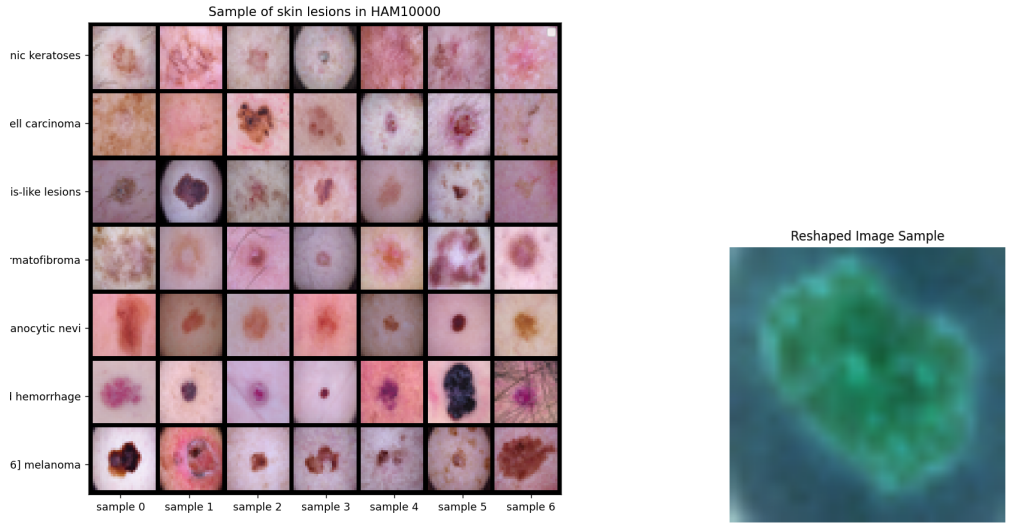
Sustainability in research also extends to the longevity and reproducibility of our work. By using established, public datasets and transparent methodologies, we aim to enable other researchers to replicate and build upon our findings, contributing to sustainable progress in the field. This approach ensures that our research not only addresses immediate scientific questions but also serves as a valuable resource for future studies, aligning with broader goals of sustainability in scientific inquiry. The code is available here: github.com/Tobias-Johannesson/II2202-DataAugmentation-GANs. The hope is that experts in the field can further improve the models and methods, tweak the hyperparameters, architectures, or methods to either confirm the conclusion with more certainty, or find a new conclusion that can help the community of the data science domain.

3.3 Experimental Setup

In our experimental setup, we will use the HAM10000 dataset and train the VGG model using a combination of real and synthetic data generated by our set of data augmentation techniques. The use of AUC score, f1-score, recall, and accuracy as our evaluation metrics will allow us to assess the effectiveness of different data augmentation techniques comprehensively. By varying sample sizes the findings can potentially be applied to a wider range of data sets and generalize better, but still does not require a new set of data exploration and pre-processing. Using multiple runs, and taking the average will provide more reliable results by minimizing the risk of working with local optima or minima. This approach not only leverages the strengths of the methodologies in the cited papers but also addresses their respective limitations, thus providing a robust and scalable solution to the challenges of data augmentation in high-dimensional image datasets. By adopting this integrated methodology, we aim to answer our research question effectively: How does combining enhanced GAN methodologies with other data augmentation techniques affect the performance of models on complex datasets? Our approach is designed to generate diverse, high-quality synthetic data and evaluate its utility in enhancing the performance of deep learning models which is our downstream task of choice. Thereby contributing valuable insights to the field of data science and machine learning.

3.3.1 The Dataset and Model

The HAM10000 dataset, a collection of dermatoscopic images of skin lesions, is used throughout the study. The HAM10000 contains 10015 skin lesion images represented in RGB color channels, and out of them, 6705 are melanocytic nevi (NV); 1113 are melanoma (MEL); 1099 are benign keratosislike lesion (BKL); 514 are basal cell carcinoma (BCC); 327 are Actinic keratosis/Bowen's diseases (AKIEC); 142 are vascular(VASC) and 115 are dermatofibroma (DF)[27, 28]. The classes are highly unbalanced with some classes such as melanocytic nevi holding over 40 times more samples compared to other classes such as dermatofibroma, and vascular. Hence it is a strong example of high-dimensional unbalanced data. The dataset's 28x28 pixeled CSV version is accessible from Kaggle.



((a)) ((b))

Figure 2: HAM10000 data samples
a) Seven random samples of each class (b) Random sample after being resized to 254x254 pixels

Type of skin cancer	Sample size
melanocytic nevi (NV)	6705
melanoma (MEL)	1113
benign keratosislike lesion (BKL)	1099
basal cell carcinoma (BCC)	514
Actinic keratosis/Bowen's diseases (AKIEC)	327
vascular (VASC)	142
dermatofibroma (DF)	115
Total	10015

Table 1: HAM10000 dataset

The data is provided as individual rows, one for each image. There are 2353 columns, where the first 2352 columns represent the 784 pixels over the three color channels. The final column is the class label as an integer which can be translated into one of different cancer classifications using metadata. After dropping the label, arrays shaped **(10015, 2352)** and **(10015,)** are obtained. The data is reshaped to first (3, 28, 28) for (color channels, rows and columns of pixels). This is the standard used by PyTorch's framework for visual data. Figure 3(a) shows 7 randoms samples across all 7 classes displayed as 28x28 RGB images. The 28x28 pixeled image data is then resized to 224x224 to fit the architecture of the VGG19 model. The modified dataset is now in the following shapes **(10015, 3, 224, 224)** and **(10015)**. The modified data appears less pixeled, but with a change in quality and coloration as seen in Figure 3(b).

For our model we use the deep learning model called VGG19, and this choice is inspired by previous work such as [29] or [30] which examines how the VGG architecture can be used to predict the whether a patient has cancer or not. PyTorch is used for implementing the deep learning models, and the architectures used is VGG19 and the GANs which are further mentioned in the section on data augmentation and preprocessing. The VGG architecture allows us to build on previous research that has proven this specific architecture working well in terms of accuracy on this specific dataset [29]. The pre-trained version of VGG will be downloaded from PyTorch, and this model is trained on the generic image dataset called IMAGENET1K which can be further examined at for example, /hrefhttps://huggingface.co/datasets/imagenet-1kHuggingFace. The use of pre-trained models has proven to

minimize required training and we can benefit from their ability to detect high-level features from previous training [31].

3.3.2 Hardware and Software

Software: The training and evaluation was done using the following software and versions: Operating System: Windows 10, 64 bit updated November 2023.

kaggle: 1.5.16 keras: 2.11.0 matplotlib: 3.5.1 numpy: 1.18.5 pandas: 1.3.5 python: 3.7.7 tensorflow: 2.3.0 torch: 2.7.1 torchvision: 0.14.1 sklearn: 1.3.2 nvcc (CUDA Toolkit): 12.3

Hardware: The experiment is run on a desktop computer (which was restarted prior to starting the experiment) from 2016 with the following components: Processor: 8-cores of Intel i7-6700K, 4.00GHz. GPU: NVIDIA GeForce GTX 980 Ti Memory: 16GB of RAM at 3.2GHZ Disk: Samsung SSD 850 EVO 500GB

3.3.3 Training and Evaluation

The entire process starts with the raw dataset downloaded from kaggle, and the data will be processed to required model sizes, split into training and test sets. For both the VGG architecture and GAN data pre-processing is necessary since the date we download 28x28 pixels and the input and output sizes of our different models vary. The data is scaled up to 224x224 images to fit the VGG model and reshaped as needed to work with both PyTorch and Tensorflow which represent rgb images as (channels, width, height) and (width, height, channels) respectively. A smaller random sub-sample without replacement were taken from the data set since the entire dataset would not fit on the local machine, and to also speed up the training. Once split, the data augmentation using the GANs, SMOTE, and other techniques are applied in parallel and their generated data is combined with the original real data. Using this new combined dataset which now consists of the original training data and generated samples (from the GAN, SMOTE, and other techniques) for the minority classes, the VGG model will be fine-tuned. Once trained, different metrics such as accuracy, f1-score, recall, and AUC will be tracked using the SciKit library. All codes will be run on a local machine, which is the desktop mentioned above since the model and data require a decent chunk of memory which was not possible to set up using the free versions of cloud computing readily available. The evaluation will lastly be done on the test performance using a test split of 20% which has not been touched during any of the steps up to that point. All these steps will be built using PyTorch's and/or TensorFlow's libraries for image classification and neural networks on the hardware and software defined in the section called "Experimental Setup".

4 Results and Analysis

In the results section, we analyze and compare the model's performance across different data augmentation scenarios. Five distinct tables provide a breakdown of key predictive metrics – AUC score, accuracy, recall, and F1-Score – under various sample sizes for differing data augmentation methods. Table 2 establishes a baseline with real data, while Tables 3 to 6 explore the effects of SMOTE, GAN, their combination, and the use of two different GAN models, respectively. This comparative analysis allows us to assess the efficacy of each data augmentation technique in improving model performance, especially in terms of prediction accuracy and reliability. The diversity in results across these tables highlights the strengths and limitations of each augmentation method, offering valuable insights into their practical application in predictive modeling. For all tables including results from data augmentation the dataset is rebalanced to have a rough ratio of 1:1 amongst the classes, where instances for all classes are generated until matching the maximum number of instances for any of the classes prior to augmentation [32].

4.1 Results

This table presents the performance metrics of the model trained on real data. Metrics include AUC score, accuracy, recall, and F1-Score, providing a baseline for comparison with augmented data sets.

Table 2: VGG Only Performance Metrics

Sample Size	AUC Score	Accuracy	Recall	F1-Score
200	0.79	0.91	0.19	0.18
500	0.76	0.89	0.14	0.11
1000	0.80	0.91	0.13	0.15
1500	0.69	0.90	0.16	0.14
2000	0.82	0.91	0.14	0.12
4000	0.86	0.94	0.20	0.18

Table 3 shows the results of the model trained on data augmented with SMOTE. It includes key metrics such as AUC score, accuracy, recall, and F1-Score, highlighting the impact of SMOTE on model performance.

Table 3: SMOTE Balanced Data Performance Metrics

Sample Size	Generated Samples	AUC Score	Accuracy	Recall	F1-Score
500	1483	0.47	0.90	0.14	0.11
1000	1490	0.61	0.89	0.14	0.12
2000	1539	0.62	0.91	0.14	0.12
4000	1160	0.72	0.90	0.13	0.11

In Table 4, the performance of the model trained on GAN-augmented data is detailed. The table covers AUC score, accuracy, recall, and F1-Score, offering insights into the effectiveness of GAN for data augmentation.

Table 4: GAN Balanced Data Performance Metrics

Sample Size	Generated Samples	AUC Score	Accuracy	Recall	F1-Score
2000	1537	0.77	0.84	0.15	0.14
4000	1152	0.83	0.87	0.12	0.11

This table illustrates the model's performance when trained on a dataset augmented with a combination of SMOTE and GAN. Metrics reported include AUC score, accuracy, recall, and F1-Score, reflecting the synergistic effect of using both augmentation techniques. Here the GAN first generates half of the required samples, and the SMOTE generates the other half of the required samples to balanced out the dataset.

Table 5: SMOTE and GAN Ensemble

Sample Size	Generated Samples	AUC Score	Accuracy	Recall	F1-Score
2000	1528	0.77	0.84	0.18	0.19
4000	1243	0.83	0.87	0.17	0.18

Table 6 compares the performance metrics, including AUC score, accuracy, recall, and F1-Score, for the model trained on data augmented using two different GAN models, showcasing the variability and potential of GAN-based augmentation approaches. Each GAN generates half of the required samples to fully balance out the dataset. Table 7 is an extension using three GAN models together, and table 8 showcases five GANs.

Table 6: Two GAN Ensemble

Sample Size	Generated Samples	AUC Score	Accuracy	Recall	F1-Score
2000	1537	0.59	0.89	0.12	0.18
4000	1152	0.67	0.90	0.15	0.16

Table 7: Three GAN Ensemble

Sample Size	Generated Samples	AUC Score	Accuracy	Recall	F1-Score
2000	1491	0.57	0.88	0.14	0.17
4000	1288	0.63	0.88	0.15	0.16

Table 8: Five GAN Ensemble

Sample Size	Generated Samples	AUC Score	Accuracy	Recall	F1-Score
2000	1511	0.54	0.81	0.13	0.18
4000	1187	0.59	0.85	0.15	0.17

4.2 Discussion and analysis

Challenge in Training GANs: A significant observation from the experiment is the inherent difficulty in training GANs, particularly due to their high demand for image data. This requirement poses a substantial challenge, as it necessitates extensive computational resources and effort. The results suggest that the additional computational burden and complexity introduced by GANs may not be justifiable, especially considering the limited improvement in performance they offer compared to the use of real data. This finding aligns with existing literature that highlights the intricate balance required in training GANs, where both the generator and discriminator need to be trained simultaneously and effectively. With a decrease in AUC score upon using more GAN models it could be assumed that the GAN is not optimally set up and that the these models still need some configuration, or the technique used did not work for this use case. The images being generated might need to be manually analyzed further, but since this task is rather subjective and could a paper on its own this is left to further studies.

Impact of Data Quality on VGG Training: The experiment underscores the importance of data quality in training deep learning models. It was observed that the VGG model's performance improved significantly with an increase in the quantity of real data. This improvement leads to an assumption that high-quality synthetic data could potentially yield similar benefits. However, defining what constitutes 'good' synthetic data remains a complex challenge. Good synthetic data, in this context, would ideally be that which closely mimics the characteristics of real data, contributing positively to the model's ability to learn generalized features rather than overfitting to noise or artefacts present in poorer quality synthetic data. It might have been better to not aim for a fully balanced data set, but instead try to make it more balanced and this can be seen by the fact that a lot of new instances were generated in comparison to the sample of real data we had for example, on having 500 samples a total of 1983 were used to train the VGG model which might explain the low AUC-score of 0.47.

GAN's Discriminator Strength and Its Impact: A notable aspect of the GAN implementation in this study was the dominance of the discriminator over the generator. This imbalance often led to the generation of poor synthetic images, which may not contribute effectively to the model training. The strong discriminator, while performing well in its task of distinguishing real from fake, inadvertently hampers the generator's ability to produce diverse and realistic images. This imbalance could be a reason why the synthetic images generated by the GAN were less effective in enhancing the VGG model's performance. These findings are significant as they point to the delicate nature of GAN training, where an overly efficient discriminator can lead to poor generator performance, ultimately affecting the quality of synthetic data produced.

Our study underscores the value of real data in enhancing model performance, while also revealing that the integration of multiple complex models does not necessarily guarantee improved results. This

finding highlights the robustness of our model in adapting to varied class distributions within the dataset. Furthermore, our research emphasizes the importance of employing diverse evaluation metrics. Relying solely on accuracy would have painted an incomplete picture, potentially obscuring critical insights into the model's predictive capabilities. As we move towards the conclusion, it becomes evident that the choices in data augmentation techniques and evaluation metrics play a pivotal role in determining the effectiveness and applicability of machine learning models. These insights contribute to the current understanding of model performance optimization.

5 Conclusions and Future Work

In conclusion, it is evident that while Generative Adversarial Networks (GANs) could present a promising avenue for data augmentation, their utility is significantly influenced by the quality of the synthetic data produced and the training equilibrium between the discriminator and generator. Since the use of a CGAN was implemented over a standard GAN our results might not fully generalize to the standard GAN architecture, but the investigation highlights that training GANs for data augmentation is difficult, and offer marginal improvements at best, often at the expense of considerable computational resources and complex model setup. This underscores the necessity of optimizing GAN architectures and training methodologies to make them more effective and efficient for practical applications. On one side the data acquired here is not enough to clearly state in what conditions GANs (if any) could be helpful in creating new synthetic data samples, but it does show that the volume of real data has a correlation to stronger predictive models. So if a set of GANs could create new data with enough resemblance to the real data this could potentially improve the downstream tasks, but from the data found here this was not the case. There is a need to analyze the data being generated by the GAN models and ensembles to more clearly state the effect GANs could have on the predictive performance of a model such as the one used in this paper.

Our experiments also reveal the importance of model choice in data processing. The efficacy of simpler models, as opposed to larger, more complex ones, aligns with the principle of Occam's razor, suggesting that straightforward solutions often yield comparable, if not superior, results. This is particularly relevant in scenarios where computational resources are limited, as seen in our inability to fully exploit large GANs and extensive datasets. Moreover, the study indicates that while a balanced datasets can impact performance, adjusting the loss function might sometimes offer a more cost-effective alternative. By just using the pre-trained VGG-model, the predictive performance was stronger and required less than half the computational resources needed compared to both training the GAN and VGG.

The exploration conducted in this study brings us to the conclusion that although augmented data can enhance model training, the true potential lies in the strength of the model and the quality of data preprocessing. Our results do not conclusively demonstrate that augmented data significantly improves accuracy, signaling a need for further investigation in this area. Future research should therefore focus on refining data augmentation models to more closely mimic real data and balance class distributions. This means that first a probably good set of GAN models could be used to generate data, and from there conduct this experiment again. Additionally, there is a pressing need for developing smaller, more efficient models that can deliver high performance on complex datasets without necessitating extensive computational power. The goal should be to find a set of smaller or more efficient models, pre-processing, and/or data augmentation techniques, making it more accessible and efficient in diverse data environments. One final interesting idea that could be studied further is to first train an initially strong model on the original dataset, and then generate artificial data which we test on the that model. Then from the artificial samples that provide the wrong classification we could throw these away, and thereby hopefully only have samples with a set of properties that resembles the original data. This would of course require that the artificially generated samples are generated in a way that captures some useful aspects to begin with so that these samples are not pure noise which would defeat the purpose of testing a generated samples usefulness.

In the initial scope of this study, we intended to explore Wasserstein Generative Adversarial Networks (WGANs) alongside GANs and SMOTE. However, due to time constraints and the complexity involved

in effectively implementing and understanding WGANs, this aspect could not be included in the current research. Future work could beneficially extend into this area, leveraging the foundational insights gained here to explore the potential advantages of WGANs in data augmentation for high-dimensional, imbalanced datasets.

Ultimately, this research contributes to the ongoing discourse in the field of machine learning and data science, particularly in the context of data augmentation. While it presents certain limitations, it also lays the groundwork for future studies aimed at enhancing the utility and efficiency of data augmentation techniques in high-dimensional, imbalanced data scenarios.

References

[1] Y. Zhai, Y.-S. Ong, and I. W. Tsang, "The emerging "big dimensionality"," *IEEE Computational Intelligence Magazine*, vol. 9, no. 3, pp. 14–26, 2014. [Online]. Available: https://doi.org/10.1109/MCI.2014.2326099

[2] H. Özköse, E. S. Arı, and C. Gencer, "Yesterday, today and tomorrow of big data," *Procedia-Social and Behavioral Sciences*, vol. 195, pp. 1042–1050, 2015. [Online]. Available: https://doi.org/10.1016/j.sbspro.2015.06.147

[3] D. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1–32, 01 2000. [Online]. Available: https://www.researchgate.net/publication/220049061_High-Dimensional_Data_Analysis_The_Curses_and_Blessings_of_Dimensionality

[4] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, "Digital medicine and the curse of dimensionality," *NPJ digital medicine*, vol. 4, no. 1, p. 153, 2021. [Online]. Available: https://doi.org/10.1038/s41746-021-00521-5

[5] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin cancer classification with deep learning: a systematic review," *Frontiers in Oncology*, vol. 12, p. 893972, 2022. [Online]. Available: https://doi.org/10.3389/fonc.2022.893972

[6] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," pp. 303–311, 2018. [Online]. Available: https://doi.org/10.1007/978-3-030-01201-4_33

[7] M. A. H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," *Decision Support Systems*, vol. 53, no. 1, pp. 226–233, 2012. [Online]. Available: https://doi.org/10.1016/j.dss.2012.01.016

[8] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*. Springer, 2004, pp. 39–50. [Online]. Available: https://doi.org/10.1007/978-3-540-30115-8_7

[9] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022. [Online]. Available: https://doi.org/10.1016/j.gltp.2022.04.020

[10] E. Strelcenia and S. Prakoonwit, "A survey on gan techniques for data augmentation to address the imbalanced data issues in credit card fraud detection," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 304–329, 2023. [Online]. Available: https://doi.org/10.3390/make5010019

[11] Y. Zhang, Z. Wang, Z. Zhang, J. Liu, Y. Feng, L. Wee, A. Dekker, Q. Chen, and A. Traverso, "Gan-based one dimensional medical data augmentation," *Soft Computing*, pp. 1–11, 2023. [Online]. Available: https://doi.org/10.1007/s00500-023-08345-z

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[13] D. Saxena and J. Cao, "Generative adversarial networks (gans) challenges, solutions, and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–42, 2021. [Online]. Available: https://doi.org/10.1145/3446374

[14] Google, "Overview of gan structure," accessed: 2024-01-13. [Online]. Available: https://developers.google.com/machine-learning/gan/gan_structure

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1710.10196

[16] G. Ramponi, P. Protopapas, M. Brambilla, and R. Janssen, "T-cgan: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling," *arXiv preprint arXiv:1811.08295*, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1811.08295

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[18] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64 606–64 628, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3074243

[19] O. O. Abayomi-Alli, R. Damasevicius, S. Misra, R. Maskeliunas, and A. Abayomi-Alli, "Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 8, pp. 2600–2614, 2021. [Online]. Available: https://doi.org/10.3906/elk-2101-133

[20] R. Polikar, "Ensemble learning," *Ensemble machine learning: Methods and applications*, pp. 1–34, 2012. [Online]. Available: https://doi.org/10.1007/978-1-4419-9326-7_1

[21] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022. [Online]. Available: https://doi.org/10.1016/j.engappai.2022.105151

[22] Y. Wen, G. Jerfel, R. Muller, M. W. Dusenberry, J. Snoek, B. Lakshminarayanan, and D. Tran, "Combining ensembles and data augmentation can harm your calibration," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2010.09875

[23] G. Eilertsen, A. Tsirikoglou, C. Lundström, and J. Unger, "Ensembles of gans for synthetic training data generation," *arXiv preprint arXiv:2104.11797*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2104.11797

[24] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1511.01844

[25] J. Cui, L. Zong, J. Xie, and M. Tang, "A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data," *Applied Intelligence*, vol. 53, no. 1, pp. 272–288, 2023. [Online]. Available: https://doi.org/10.1007/s10489-022-03361-2

[26] L. Luzi, R. Balestriero, and R. G. Baraniuk, "Ensembles of generative adversarial networks for disconnected data," *arXiv preprint arXiv:2006.14600*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.14600

[27] K. H. Tschandl P, Rosendahl C, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci Data*, vol. 5, 2018. [Online]. Available: https://doi.org/10.1038/sdata.2018.161

[28] P. Tschandl, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018. doi: 10.7910/DVN/DBW86T. [Online]. Available: https://doi.org/10.7910/DVN/DBW86T

[29] N. Abuared, A. Panthakkan, M. Al-Saad, S. A. Amin, and W. Mansoor, "Skin cancer classification model based on vgg 19 and transfer learning," in *2020 3rd International Conference on Signal Processing and Information Security (ICSPIS)*. IEEE, 2020, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICSPIS51252.2020.9340143

[30] M. A. A. Milton, "Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge," *arXiv preprint arXiv:1901.10802*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1901.10802

[31] M. Subramanian, K. Shanmugavadivel, and P. Nandhini, "On fine-tuning deep learning models using transfer learning and hyper-parameters optimization for disease identification in maize leaves," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13 951–13 968, 2022. [Online]. Available: https://doi.org/10.1007/s00521-022-07246-w

[32] Y.-C. Wang and C.-H. Cheng, "A multiple combined method for rebalancing medical data with class imbalances," *Computers in Biology and Medicine*, vol. 134, p. 104527, 2021. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.104527