Computational Creativity: Artificial Intelligence Models for the Generation and Classification of Affective Artwork and their Human Evaluation

A Thesis Report

submitted by

YASH RAJ

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

Thursday 29th June, 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Computational Creativity: Artificial Intelligence Models for the Generation and Classification of Affective Artwork and their Human Evaluation**, submitted by **Yash Raj**, to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Ganesh Bagler
Thesis Supervisor
Professor
Infosys Center for Artificial Intelligence
IIIT Delhi - 110020

Place: New Delhi
Date: 29 June 2023

ACKNOWLEDGEMENTS

I would like to extend my sincerest appreciation to my advisor, Prof. Ganesh Bagler, who has been an invaluable guide throughout this project. His constant guidance and support have enabled me to navigate every obstacle encountered in this project. I have learned and grown immensely through his constructive feedback on my work. His invaluable insights and expertise have been instrumental in shaping my research, and I am deeply grateful for his mentorship.

I would also like to acknowledge my family and friends for being a consistent source of support through exceptionally trying times in my life. I express my gratitude to Dr. Arjun Ray for providing access to his GPU servers.

I am grateful for my time at IIIT-Delhi, which has proved to be a turning point in my life and has provided me with the opportunity to learn. I deeply appreciate the faculty and staff at IIIT-Delhi for their dedication and commitment to providing the best education for their students.

Lastly, I would like to express my heartfelt gratitude to all of those who have supported me throughout this journey, without whom this achievement would not have been possible.

ABSTRACT

Visual artwork is among the most salient forms of human expression. From prehistoric cave paintings to Renaissance and modern art, paintings have been a powerful medium for expressing emotions. With the advent of computing and artificial intelligence, visual arts may no longer be exclusive to human creativity. Computational creativity involves the study of creative endeavors ranging from creative writing, poetry, painting, music, and science to sports through computational approaches. On the intersection of art and computer science, this thesis involves implementing artificial intelligence models to generate and classify affective artwork and their human evaluation.

Rooted in the WikiArt data of over 80,000 paintings and their emotion labels from ArtEmis, we implement Generative Adversarial Networks for generating paintings with desired emotional content. We first experiment with two broad classes of emotions (positive and negative) to further deal with nuanced affective categories, viz. amusement, awe, contentment, excitement, anger, disgust, fear, and sadness. Besides computationally generating affective artwork, we also implement classification models and validate their performance using relevant metrics. Projected GANs, StyleGAN2-ADA, and StyleGAN3 are employed for generating artwork for binary and multi-class models to achieve an FID score of 7.84 for the StyleGAN2-ADA architecture. ResNet50-V2 presents the highest accuracy for the binary classification experiment at 72%.

Beyond the computational evaluation of the generated artwork, we created a 'Turing Test for Artist.' This test randomly presents images of human artwork, and those made with artificial intelligence to a human evaluator and registers their assessment. For every image, the test also records the binary human assessment of the affective content of the artwork. We assess the quality of the generation and classification models after conducting the Turing Test with a sizeable number of evaluators. We conclude that while the artificial intelligence approach is capable of producing affective artworks that compete with human creativity, it is far from replacing it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CNN	Convolutional Neural Networks
DL	Deep Learning
GAN	Generative Adversarial Networks
DNN	Deep Neural Networks
ML	Machine Learning
NN	Neural Networks
RNN	Recurrent Neural Networks
FID	Fréchet Inception Distance
KID	Kernel Inception Distance
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
MCC	Matthew's Correlation Coefficient
ROC Curve	Receiver Operating Characteristic Curve
AUC	Area Under the Curve
ROC-AUC	Area Under the ROC Curve

CHAPTER 1

Introduction

Throughout history, humans have used various art forms such as painting, sculpture, literature, music, and dance to express their emotions and ideas [1]. Art has also been used to reflect and comment on society, politics, and cultural norms. It is essential in contemporary culture and societal issues [2]. Various studies have highlighted the role of art as a powerful mechanism for evoking an emotional response [3, 4].

For decades after their invention first artificial neural network in the 1950s, computational models have been underutilized to explore complex human endeavors mainly due to limited processing capabilities. With increasing computational abilities, computational approaches were used for performing a broader range of tasks, from simple calculations to complex tasks such as emulating human competence for video games [5]. In the recent past, generative algorithms such as Generative Adversarial Networks (GANs), Diffusion Models [6, 7, 8], and Variational Autoencoders (VAE) [6, 9, 10] have emerged as powerful tools for generating visual artworks.

Generative art aims to create unique and exciting visual outputs that the artist does not predetermine. Artists and painters use their creative abilities to produce visual artworks through paintings to express their thoughts, ideas, cultural influences, and emotions [11]. The advent of algorithms capable of generating images, texts, and videos has created exciting propositions. 'Can computers generate visual artworks with the same emotional depth and style as human-created art?' 'Can computers achieve match human creativity?'

In this research, we primarily investigate the capabilities of GANs to generate affective artwork evoking emotional responses resembling those produced by human creations. We explore how machine learning models can effectively capture emotional attributes in the generated artwork by implementing GANs to generate affecive paintings and CNN models for their classification. With these studies, we aim to advance the understanding of the potential of GANs as a tool for creating visually appealing and emotionally expressive paintings. We use CNN models to classify artwork based on their emotional

attributes. Finally, we have implemented a Turing Test framework for human evaluation of generated artwork to assess the performance of generative and classification models.

1.1 Generative Adversarial Networks (GANs)

The notion of Generative Adversarial Networks (GANs) was introduced in 2014 by Goodfellow et al. [12]. GANs comprise two neural network components: the Discriminator model (Discriminator) and the Generative model (Generator). These models are trained simultaneously, with Generator learning to imitate the training data distribution. At the same time, Discriminator decides the likelihood of a sample belonging to the training data rather than the generated data. Training of the Generator aims to maximize the chances of the Discriminator making errors. The GAN framework employs an adversarial approach, resembling a competitive game between two players. In this adversarial game, the discriminator learns to accurately distinguish between the real and generated instances, while the generator aims to minimize the discriminator's ability to classify. To illustrate this concept further, imagine a team of forgers (Generators) creating fake currency and attempting to bypass detection while the police (Discriminator) strive to catch them. Through this competition, both teams are motivated to enhance their skills until the fake money becomes virtually unnoticeable from genuine currency.

1.1.1 Conditional GANs

Mirza et al. [13] introduced a method to enhance the control over data generation in vanilla GANs by incorporating conditioning. In conditional GANs, additional information is provided to the unconditioned GAN, typically a class label. This auxiliary information enables the unconditioned GAN to generate data that aligns with the specific class label.

1.1.2 StyleGANs

Developed by NVIDIA, Style Generative Adversarial Networks (StyleGANs) [14] has emerged as a powerful deep learning model. These networks build upon the origi-

nal GAN algorithm by incorporating insights from style transfer research, leading to significant advancements in image generation. NVIDIA further refined StyleGAN by introducing StyleGAN2 [15] in 2019 to incorporate path length regularization and image mixing techniques. StyleGAN2-ADA [16] extended this progress by introducing the Adaptive Discriminator Augmentation (ADA) training method, allowing the discriminator to adapt and handle diverse generated images, even with limited datasets. Introduced in 2021, the most recent addition to the StyleGAN series is StyleGAN3 [17]. Despite exhibiting comparable performance to StyleGAN2, StyleGAN3 demonstrates complete invariance to translation and rotation, even at subpixel levels. This characteristic makes StyleGAN3 particularly well-suited for applications in video and animation, opening new possibilities for creating realistic and dynamic visual content. The evaluation of StyleGANs commonly relies on the Fréchet Inception Distance (FID) score [18], which quantifies the similarity and quality of the generated images compared to real ones.

1.1.3 Projected GANs

Based on the discovery that the discriminator cannot fully utilize features from deeper layers of the pretrained model. Sauer et. al. introduced a more efficient approach called Projected GAN that combines features from different channels and resolutions [19]. This variation of GAN demonstrated compatibility with high-resolution images and achieved state-of-the-art results on 22 benchmark datasets, as evaluated by the FID metric. Moreover, Projected GANs exhibit significantly faster convergence than previous methods, with a convergence speed up to 40 times faster. When applied to WikiArt painting datasets, Projected GANs were shown to achieve the lowest FID scores, surpassing the performance of StyleGAN2-ADA [16] and FastGAN [20]. With the image resolutions of 256×256 and 1024×1024 datasets (WikiArt data), the FID scores of 27 and 32 were reported, respectively.

1.1.4 Emotion-based Art Generation

Among the few studies that probed the affective attributes of paintings is a research article titled ‘Art Creation with Multi-Conditional StyleGANs’ [21]. This paper imple-

mented StyleGANs to generate art that closely resembles human paintings. The authors combined ArtEmis and Wikiart datasets to create a comprehensive dataset called EnrichedArtEmis. They employed the StyleGAN2-ADA architecture that supports conditional capabilities to train a multi-conditional GAN. Due to the low resolution of the generated images, a significant amount of training data was required. Other than the emotional attribute of the artwork, the conditions considered in the model included art style, genre, painter, and other associated tags. The generated images produced by this network were assessed by the authors to be visually compelling and often indistinguishable from human-created art. The model performance was evaluated using the FID score, with an impressive score of 4.67. The study reported higher FID scores of 10.51 and 9.74 on the metric 'Emotion Intra-FID'. Other than the FID score, the paper does not provide any human validation for the quality of generated images.

Another study, 'The Emotional GAN: Priming Adversarial Generation of Art with Emotion', probed the affective aspect of paintings [22]. In this article, the authors utilized WikiArt (2017) [23] and MoMA [24] datasets. Emotions were associated with paintings using a CNN classifier trained on the human-labeled data. The AC-GAN [25] model was applied to these images to train the GAN and generate art based on emotions. The use of data from different sources with inconsistent labeling strategies is one of the shortcomings of this study. Also, machine-labeled emotion tags are a potential noise source due to the model's under-par performance. Also, the study did not provide any quantitative evaluation of model performance.

1.2 Prior Studies on Artwork Classification

Zhao et. al. presented a comprehensive evaluation of seven models across three diverse datasets to compare their performance in art classification, both with and without transfer learning [26]. The models were specifically assessed for their ability to classify genres, styles, and artists. Additionally, they investigated the challenges encountered by computers when classifying art. The study utilized three prominent painting datasets, Painting-91 [27], WikiArt-WikiPaintings [26], and MultitaskPainting100k [28], widely recognized as benchmark datasets in art classification. Notably, the models pre-trained on ImageNet [29] exhibited the most favorable outcomes in art classification. This find-

ing suggests that the classification abilities developed in real-world image classification tasks can effectively transfer to the domain of art classification. Among the models tested, ResNeSt (a variant of ResNet [30]) and EfficientNet [31] classifiers presented the best performance under different experimental conditions.

Aslan et al. introduced a new version of ArtGraph, an artistic knowledge graph for recognizing emotion in artworks [32]. The ArtEmis dataset [33] was used for training and testing purposes. They proposed an emotion classification system that integrates the knowledge graph (ArtGraph) and visual features to enhance the model's capability for recognizing emotions evoked by painting. The reported accuracy for binary emotion and multi-emotion classification on the proposed model was 81.57% and 45.39%, respectively. The study suggests a complex interconnection between style, genre, and emotion that can be harnessed for various applications including automated art analysis. Despite its merits, the model evaluation in this study solely relied on accuracy as a metric, without considering other vital metrics such as precision, recall, F1-score, and Matthew's correlation coefficient (MCC). As mentioned in Chapter 2 the ArtEmis dataset exhibits class imbalance. Therefore, it is essential to report the above-mentioned metrics for better insights into the potential biases of the model.

1.3 Motivation

While previous studies have probed various aspects of visual artwork generation, its emotional attributes and affective quality haven't received much attention.

In one study that focused on the affective quality of the artwork, a multi-conditional model was developed by incorporating emotion as one of the five conditions. However, there is a significant gap in the human evaluation of computer-generated affective artwork. Aslang et. al. identified several gaps in their survey involving emotion recognition in visual artworks [32]. The lack of a well-rounded assessment of models was notable with the sole focus on accuracy as the metric of evaluation. Also, the study did not provide a clear interpretation of their results, leaving room for further analysis and interpretation, discussed in further chapters.

Assessing and appreciating emotions is a complex task even for humans due to a complex interplay of multiple factors and subjective judgements [34]. This research aims

to generate visual artworks with desirable emotional attributes using GANs and assessment of emotion using classification models. The idea is also to assess the affective attributes of the generated artworks using, both, the classifiers and human subjects. GANs and classifier models employ distinct architectures for feature selection from the training dataset. Examining their results is an intriguing and interesting task given the disparity in their approach. Comparing and contrasting can provide valuable insights into the differences between GANs and classifiers with consequences for their performance and ability to capture relevant features. We intend to assess the results from the classifiers through human evaluation to examine the alignment between the machine label and human perceptual judgment. In addition to the conventional evaluation metrics such as FID, KID, precision, and recall, the inclusion of human assessment provides valuable insights into the disparities between the machine evaluation process and human judgment. Our study aims to bridge the gap between human and machine-generated artwork, contributing to future advancements in Computational Creativity.

CHAPTER 2

Data Compilation and Annotation

In this research, we used data of artworks and annotations describing their emotional attributes from ArtEmis [33], ArtEmis Dataset V2.0 [35], and the WikiArt dataset [23].

2.1 WikiArt

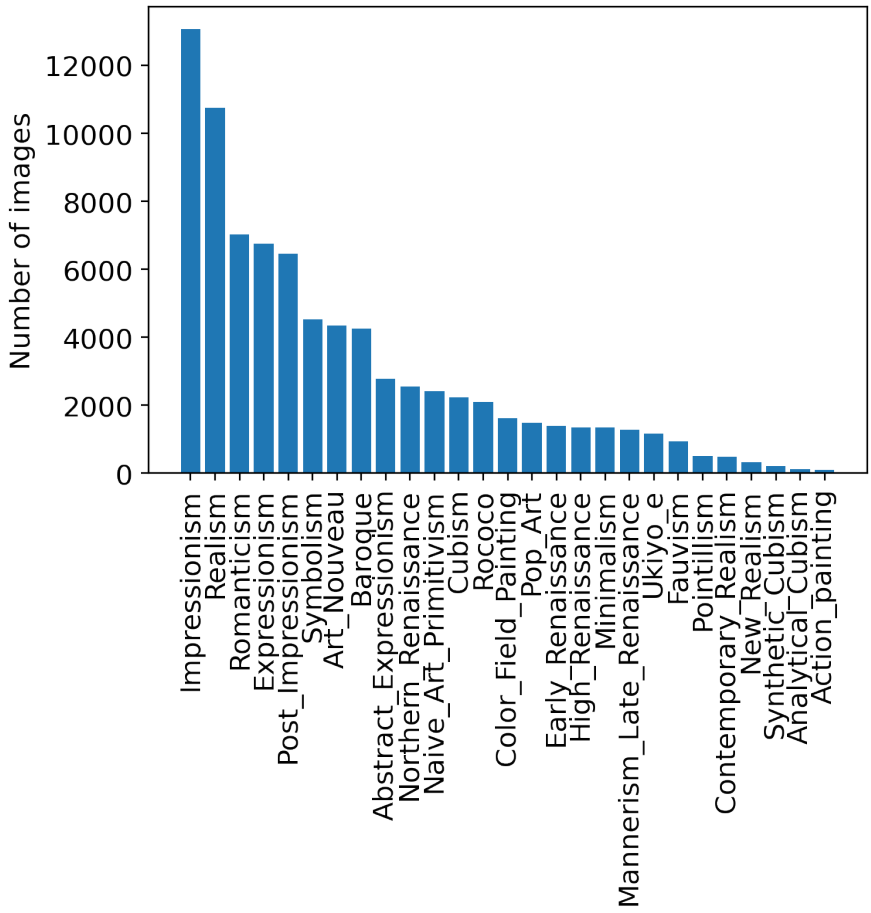


Figure 2.1: Style-wise artwork distribution in WikiArt dataset.

WikiArt Dataset [23] contains 81,444 visual artwork images with 27 different art styles and provides information about their style, genre, artist's name, and title. The data was obtained from `https://archive.org/details/wikiart-dataset` (Last Updated on 29 January 2020). Figure 2.1 presents the distribution of images across

styles. Among the most frequent images are those from the classes Impressionism, Realism, Romanticism, Expressionism, Post-Impressionism, Symbolism, Art-Nouveau, Baroque, and others.

2.2 ArtEmis

ArtEmis dataset [33] contains 454,684 emotion attributions and explanations from human volunteers for 80,031 artworks from WikiArt. All WikiArt artworks were annotated by asking at least five annotators per artwork to express their dominant emotional reactions. They are required to decide their emotional reactions by looking at the artwork and explaining their response. The annotators were asked to indicate their dominant reaction by selecting one emotion among the eight emotions (*anger*, *amusement*, *awe*, *contentment*, *disgust*, *excitement*, *fear*, and *sadness*) or as 'something-else,' the ninth option. This 'something-else' option allows the annotators to express emotions not explicitly listed and to explain why they might not have had any strong emotional reaction, such as feeling indifferent to the artwork. In all cases, after this step, the annotator was asked to provide a detailed explanation for their choice in the free text that would include specific references to visual elements in the artwork.

2.2.1 ArtEmis Statistics

Each artwork from WikiArt was annotated by a minimum of five individuals, and 701 had more than 41 annotators. Around 96% of the total artworks were annotated by either 5 or 6 annotators, as shown in Figure 2.2. On average, there were 5.68 annotators per artwork. *Contentment* was the most frequently annotated emotion, with 126,129 annotations, while *anger* was the least frequent, with only 6,640 annotations, as shown in Figure 2.3.

For the present study, the emotions were categorized into *positive* (*amusement*, *contentment*, *awe*, *excitement*) and *negative* (*anger*, *disgust*, *fear*, *sadness*) emotions, with 282,023 annotations for the positive and 119,685 annotations for the *negative* class as depicted in Figure 2.3. The *positive* labels accounted for 62% of all annotations, with 26% *negative* labels, and the rest were of the 'something-else' class.

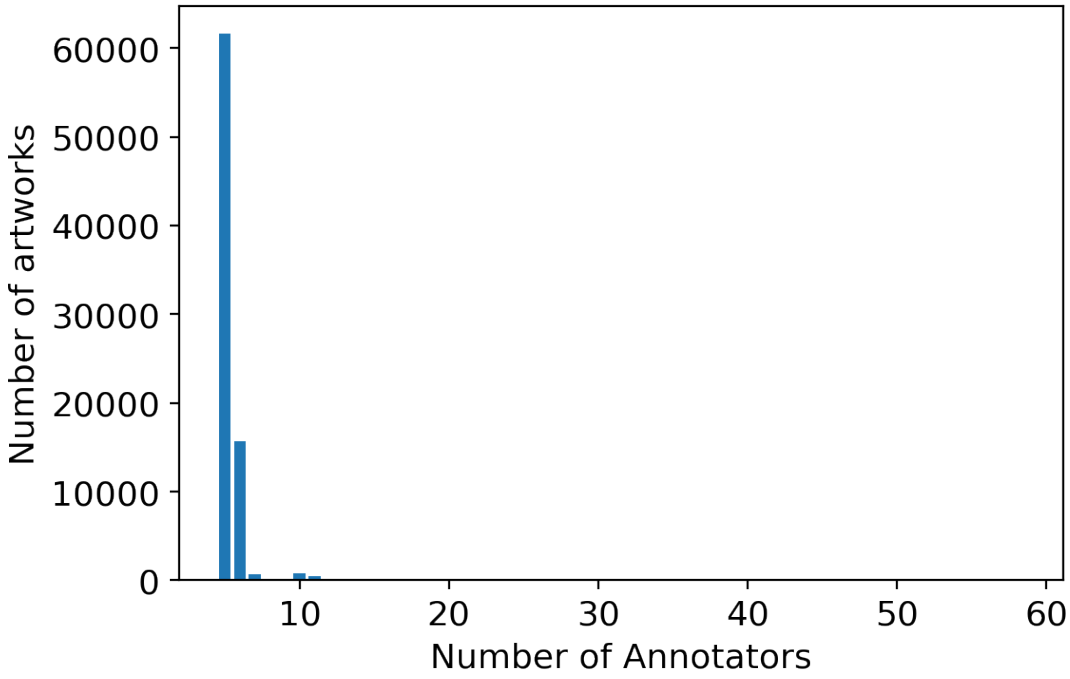


Figure 2.2: Frequency distribution for the number of annotators for ArtEmis.

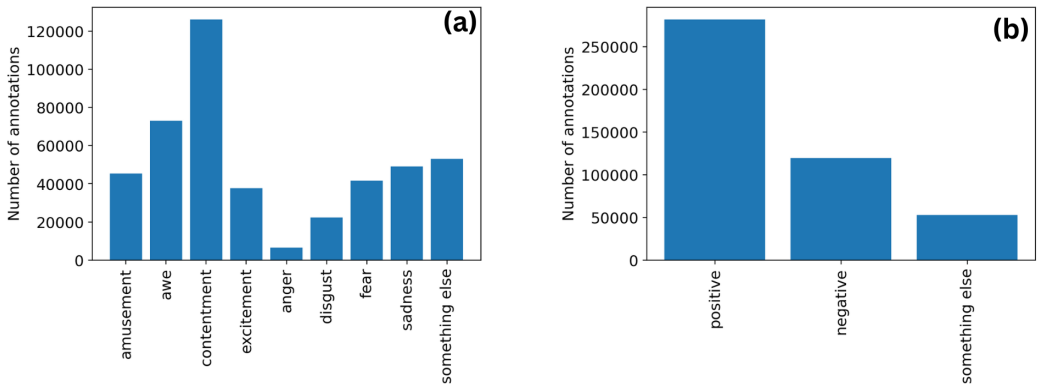


Figure 2.3: Statistics of emotion annotations in ArtEmis dataset. (a) Number of annotations for each of the nine emotion classes. (b) Number of annotations for the *positive* (*amusement*, *contentment*, *awe*, *excitement*) and *negative* (*anger*, *disgust*, *fear*, *sadness*) emotions, and those for 'something-else'.

2.3 ArtEmis v2.0

The ArtEmis v2.0 [35] is a combined dataset that represents an extended version of the original ArtEmis dataset, with an additional 260,533 annotations totaling 692,682 annotation instances. There is no significant rise in the number of artworks. There are notable emotional biases with higher instances of *contentment*, *awe*, and *sadness*. The *positive* emotions class constitutes almost 2.4 times the number of annotations in the *negative* emotions class, as shown in Figure 2.3.

2.3.1 ArtEmis V2.0 Statistics

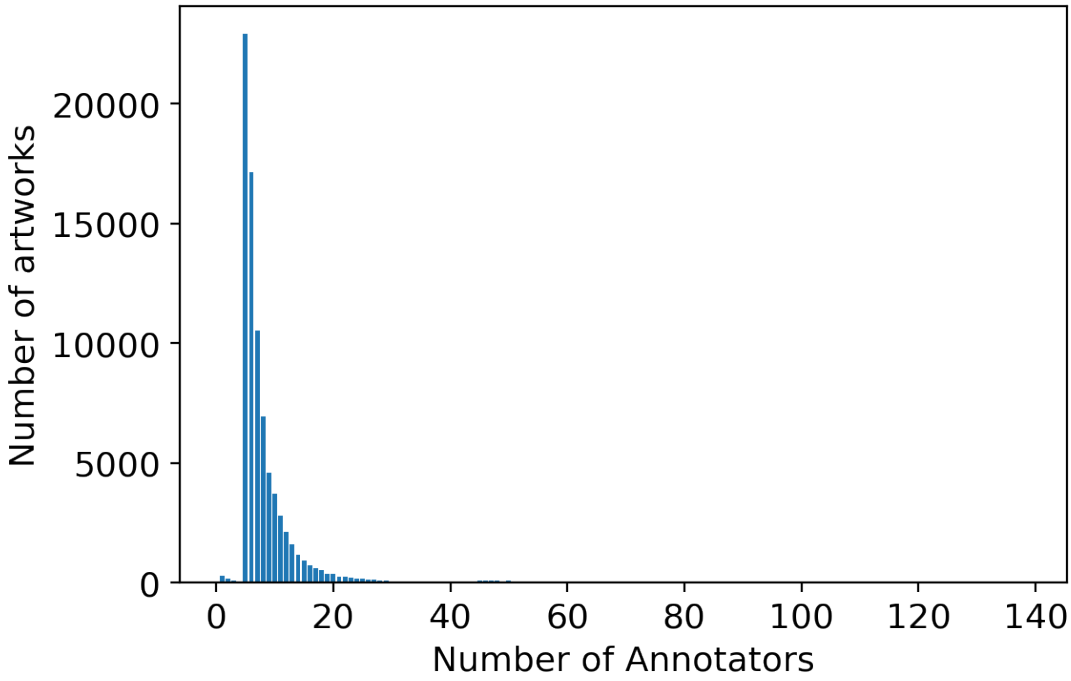


Figure 2.4: Frequency distribution for the number of annotators for ArtEmisV2.0.

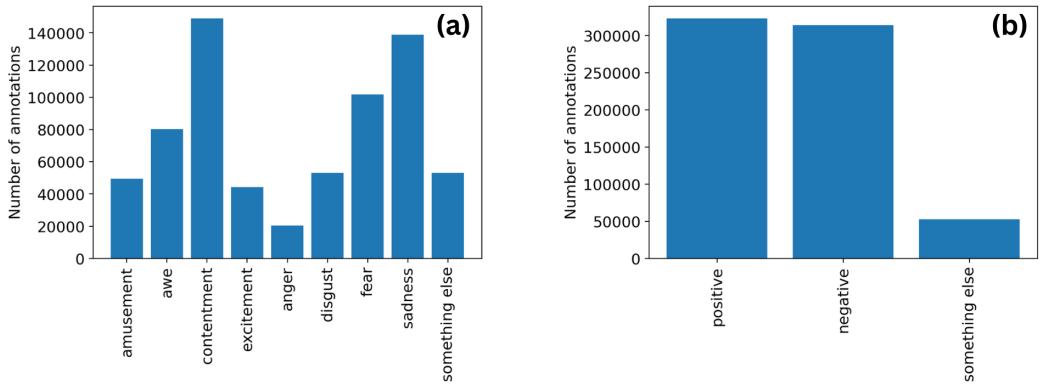


Figure 2.5: Statistics of emotion annotations in ArtEmisv2.0 dataset. (a) Number of annotations for each of the nine emotion classes. (b) Number of annotations for the *positive* (*amusement*, *contentment*, *awe*, *excitement*) and *negative* (*anger*, *disgust*, *fear*, *sadness*) emotions, and those for 'something-else'.

The combined dataset exhibits an average of 8.56 annotations per artwork. The frequency distribution for the number of annotators is shown in Figure 2.4. The new annotations were intentionally added to address the class imbalance of emotions. Most of the new annotations belong to the class of *sadness*, *fear*, and *disgust*, thereby increasing the proportion of *negative* class instances and mitigating the class imbalance in the dataset. *Positive* annotations account for 47%, while *negative* annotations constitute 43% of the total annotations shown in Figure 2.5.

CHAPTER 3

Methodology

Toward achieving the thesis objectives, we put together a computational framework (Figure 3.1) involving annotating artworks with the dominant emotion, implementing generation and classification models, and human assessment of generated images. The experiments used two broad paradigms; binary emotion class, in which the emotions were bundled into positive and negative classes, and multi-emotion class, in which each emotion was treated separately.

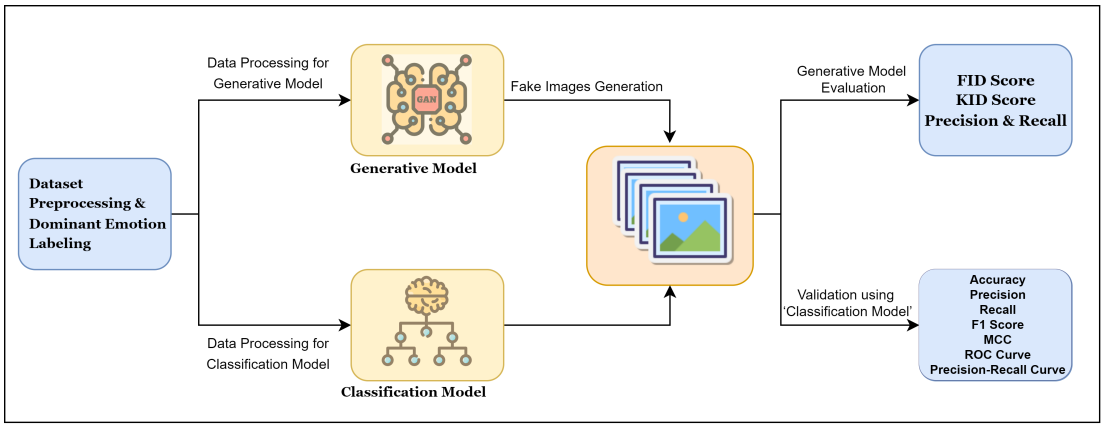


Figure 3.1: Computational Framework

3.1 Computational Framework

Figure 3.1 depicts the computational framework implemented for generating and classifying affective artworks. It comprises data collection, preprocessing, dominant emotion labeling, implementation of generation and classification models, and their subsequent evaluations.

Artworks from ArtEmis V2.0 dataset [35] were used as training data for the generation task. The preprocessing steps used to prepare the dataset are discussed in Section 3.1.1. The studies were conducted in two stages. To begin with, we treated the data in binary classes, where the emotions were categorized as *positive* or *negative*. Such coarse-graining simplified the image classification and generation problem by reducing the

number of classes. Going further, we trained GANs using multi-emotion class labels (*Amusement*, *Awe*, *Contentment*, *Fear*, and *Sadness*) for five emotion classes. The quality of the generated images was then evaluated using the Fréchet Inception Distance (FID) score [18], Kernel Inception Distance (KID) [36], precision [37], and recall [37]

Classification models were trained for binary and multi-class experiments using the ArtEmis V2.0 dataset [35], with variations in the preprocessing steps. The performance of the classification model was evaluated in terms of accuracy, precision, recall, F1-score, Matthews correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) curve, and Precision-Recall curves. These models, including GANs and classifiers, aim to capture the underlying data distribution and extract visual features such as colors, objects, and styles. The GANs and classifiers underwent further human evaluation using two types of assessments: 'The Turing Test' and 'Guess the Emotion.' The Turing Test evaluates the ability of the generative model to deceive human evaluators, whereas in the 'Guess the emotion' task, evaluators assign emotional attributes to displayed images. It is important to note that human evaluation was conducted exclusively for the binary emotion classification experiment.

3.1.1 Dominant Emotion Labeling

Tagging artworks with emotion labels in ArtEmis was done with the help of multiple annotators leading to multiple labels assigned to the same artwork. For instance, the 1876 artwork 'Sowar - The Messenger of The Government' in the Realism genre by Vasily Vereshchagin has two annotations each for *amusement*, *contentment*, and *fear*, and one annotation for *awe*. On average, each painting had 8.56 annotations, with each annotator assigning one emotion. To use these data for training image generation models and classifiers, we needed a single dominant emotion unambiguously attached to each artwork. For this purpose, we used a many-to-one function f: $X \rightarrow Y$, where X represents the set of artworks and Y represents the set of emotions. This function ensures that each artwork was tagged with a single dominant emotion. The emotion was considered as dominant as long as its frequency was beyond a cutoff threshold for a given artwork.

Figure 3.2 illustrates the preprocessing steps for the binary emotion classes. We aggre-

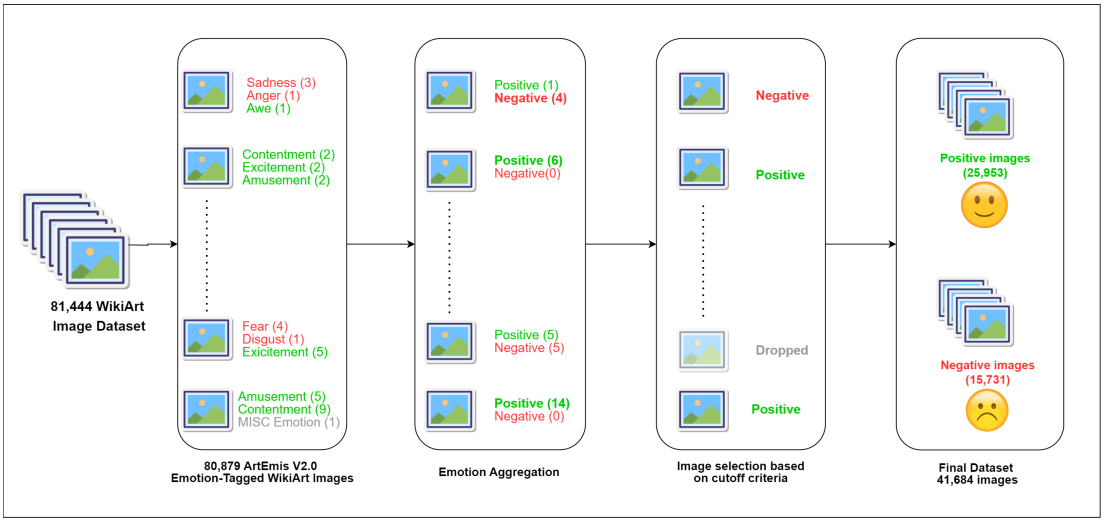


Figure 3.2: 'Dominant emotion labeling' for binary classes with positive and negative emotions.

gated the eight original emotions into *positive* and *negative* categories. For experiments with binary emotion classes, we used a cut of 66.6%, and the emotion label was decreed dominant as long as two-thirds or more of the total annotations are from that class. For example, if an artwork has seven *positive* and three *negative* annotations, it was declared as having *positive* dominant emotion. Artworks that did not have a dominant emotion were dropped out to reduce noise in the training data. Not all artworks evoke an unambiguous dominant emotion and may evoke mixed affective responses thereby introducing noise into the training dataset. Using the criterion for establishing the dominant emotion with each image, we were left with 41,684 artworks, of which 25,953 were put in the *positive* class, and the rest of 15,731 were slotted in the *negative* class.

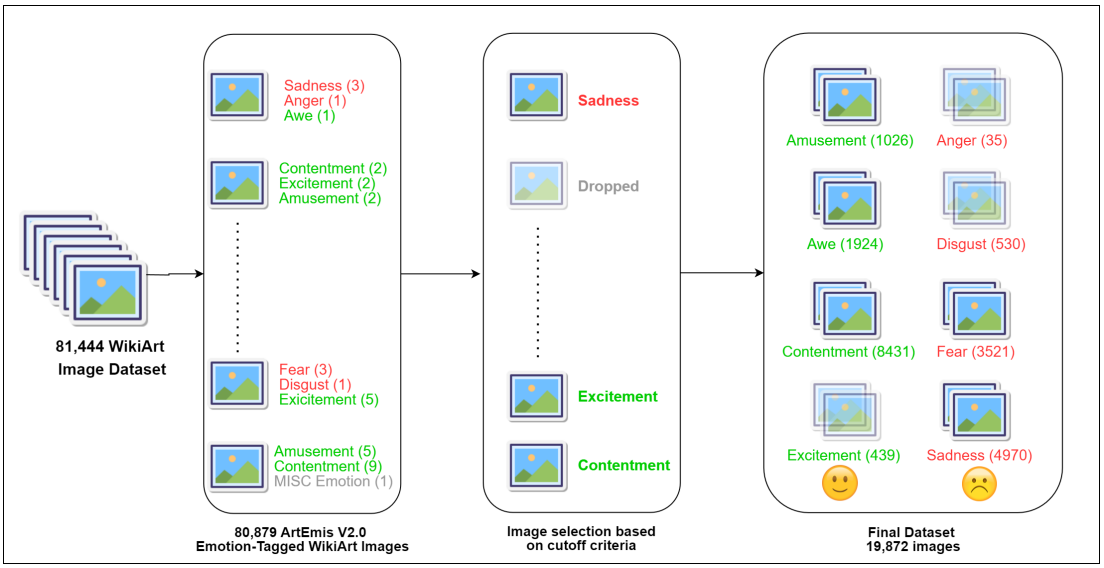


Figure 3.3: 'Dominant emotion labeling' for multi-class experiments.

Similarly, for the multi-emotion class experiment depicted in Figure 3.3, every artwork was linked with a dominant emotion if it had 50% or more annotations. The painting was dropped from the dataset otherwise to avoid images with noisy labels. We were left with a total of 19,872 artworks after dominant emotion labeling. These data had images of the following emotion classes: *Amusement* (1,026), *Awe* (1924), *Contentment* (8431), *Fear* (3,521), and *Sadness* (4,970).

3.1.2 Generative Models

In this study, we used Generative Adversarial Networks (GANs) as generative models to create artwork with desired emotional attributes. The following are the various aspects of the image generation.

1. **Dataset Preprocessing:** To meet the requirements of the generative model, the ArtEmis dataset was subjected to preprocessing as described in Section 3.1.1. The preprocessing steps involved data cleaning and normalization. All the images were downsized to 256×256 resolution.

2. **GANs:** The study utilized the Projected GANs [19] with the FastGAN Lite generator [20], StyleGAN2-ADA [16] and StyleGAN3. In, both, StyleGANs, the adaptive discriminator augmentation (ADA) technique was employed. This technique is specifically utilized when the dataset available for training is small. The ADA mechanism plays a crucial role in stabilizing the training process under limited data regimes and effectively mitigates the risk of overfitting the model [16]. All these GANs were trained on the preprocessed dataset for a certain number of 'kimg'. The number of images in thousands needed to show to the discriminator at the time of training is referred to as kimg. The hyperparameters for the GANs were set based on the 'configs.md' file provided in the GitHub repository of StyleGANs.

3. **Evaluation of Generated Images:** The quality of the generated images was evaluated using the FID score, a widely accepted metric for assessing the quality of image generation models [18]. This score measures the distance between feature vectors calculated for real and generated images. Among all the metrics computed, FID score is the most indicative of the visual quality of the generated images. A lower FID score often corresponds to better visual results [38]. Other than the FID score, we have also used the KID score, precision [37] and

recall [37]. The interpretation of precision and recall is way different when used for assessing generative models compared to their use for evaluating classification models (Figure 3.4).

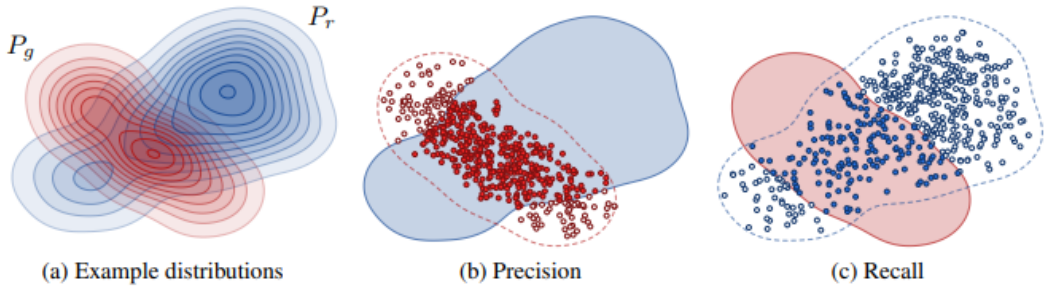


Figure 3.4: The interpretation of precision and recall for assessing the quality of generative models. (a) Probability Distributions of Real (P_r) (blue) and Generated Images (P_g) (red). (b) Precision refers to the likelihood that a randomly selected image from P_g belongs to the set of real images, P_r. (c) Recall represents the probability that a randomly selected image from P_r is included in the set of generated images, P_r.

4. **Artwork Generation:** The trained GAN models were utilized to generate artwork images. Specifically, we generated 2000 artworks per emotion class for each GAN model.

3.1.3 Classification Models

Classification is a crucial aspect of this thesis as it assesses the performance of the generative model outcomes. The following are the various aspects of the classification task.

1. **Preprocessing the Dataset**: As mentioned in Section 3.1.1, further to processing done earlier, the ArtEmis dataset was cleaned and normalized to meet the requirements of the classification algorithm. The results were found to be consistent between experiments done with two variants of image resolutions (224×224 and 256×256).

2. **Classification Algorithms**: Inception-ResNet-V2 [39], ResNet50-V2 [30], Inception-V3 [40], and VGG19 [41] classifiers were used for classification experiments. Transfer learning methods provided by the Keras library were used for efficient model training.

3. **Model Evaluations**: Metrics such as accuracy, precision, recall, MCC, F1-score, ROC, and precision-recall curve were employed to evaluate the effectiveness of the models in classification tasks. These metrics provide insights into the performance and effectiveness of the classification algorithms.

4. **Validation of Generative Models**: The images generated by the models were evaluated using the Classifier to assess their recognition and classification accuracy. This validation step helps determine the effectiveness of the generative models in producing images that the classification algorithm can label correctly.

3.1.4 Human Assessment

We created a testing framework referred as the 'Turing Test for Artist' (TTA) to evaluate the effectiveness of the generative algorithms in producing real-like images with a desired emotional response in its viewers. After showing a random image from the real or generated images, the binary classification task records the evaluator's assessment of the image as fake (computer-generated) or real artwork. For the same image, the test also registers their judgement on the emotional response as *positive* and *negative*.

This evaluation serves two primary purposes. First, to assess the quality of the generated images in deceiving human evaluators. And, second to analyze the correlation between machine and human judgments on emotional quality of the artwork. Details of the experiment design and results of TTA are provided in Chapter 5.

3.2 Experiments

Two paradigms were used when conducting experiments: binary emotion classes and multi-emotion classes. In the first paradigm, we focused on binary emotion classes, namely, *positive* and *negative* emotions. This choice was made for two reasons. Firstly, it is easier for classification models to handle a lesser number of classes. Secondly, it is also easier for humans to understand and interpret emotions when bundled into two broad categories. The GANs (Projected GAN, StyleGAN2-ADA, and StyleGAN3) and classification models were trained on the processed ArtEmis dataset comprising *positive* and *negative* images. The human assessment (TTA) was also conducted for the binary class experiment (discussed in detail in Chapter 5).

In the second paradigm, we worked with multi-emotion classes using five of the eight available emotions. Three emotion classes (*Anger*, *Disgust*, and *Excitement*) were excluded due to the limited number of associated artworks available for training the GANs. These emotions had fewer than a thousand artworks, far less than needed for effectively training a GAN. The data of artworks with the following emotions classes were used for the multi-emotion class experiments: *Amusement*, *Awe*, *Contentment*, *Fear* and *Sadness*. Similar to the earlier experiments, GANs were trained to generate 2,000 images per emotion, resulting in 10,000 generated artworks for each GAN. classification models were then trained using these generated images. A detailed explanation of these experiments can be found in Chapter 4.

CHAPTER 4

Generation and Classification of Affective Artwork

Using the processed dataset (Chapter 2) and following the described methodology (Chapter 3), below, we present the results of results and observations from generative and classification models for the binary and multi-class paradigms.

4.1 Experiments with the Binary Paradigm

4.1.1 Generative Models

Table 4.1 presents the performance evaluation of generation models implementing StyleGAN2-ADA, Projected GAN, and StyleGAN3 architectures.

Generative Model	FID Score ↓	KID Score ↓	Precision	Recall
StyleGAN2-ADA	**7.85**	**0.00156**	0.581	0.235
Projected GAN	8.42	0.00183	**0.700**	0.144
StyleGAN3	15.19	0.00565	0.568	**0.248**

Table 4.1: Performance comparison of GAN models (Binary-class).

On the FID score, Projected GAN performed better in the initial epochs but was overtaken by StyleGAN2-ADA at 2000 kimg, as shown in Figure 4.1. The FID scores for StyleGAN2-ADA, ProjectedGAN, and StyleGAN3 were 7.85, 8.42, and 15.19, respectively, with StyleGAN2-ADA achieving the best FID score. In terms of KID score, StyleGAN2-ADA performed best with a score of 0.00156, compared to that of ProjectedGAN (0.00183) and StyleGAN3 (0.00565). The comparative performance of GANs was consistent for FID and KID scores. It was observed that the model with a lower FID score also has a lower KID score and suggests better performance. The ProjectedGAN had the best precision score suggesting its superior ability to produce high-quality images similar to real ones. The recall score indicates the diversity of the generated images, and results suggest that StyleGAN3 captures most features from the training data. In the early epochs, the models had a higher precision and lower recall,

but in later epochs, the models traded precision for recall as both are necessary for effective image generation. These experiments were primarily conducted on IIIT-Delhi GPU servers with Nvidia RTX 3090 having 24 GB VRAM. With these hardware configurations, ProjectedGAN, StyleGAN2-ADA, and StyleGAN3 models progressed at 4000, 2000, and 1000 kimg per day. The fastGAN_lite generator is twice as fast as StyleGAN2-ADA and four times faster than StyleGAN3.

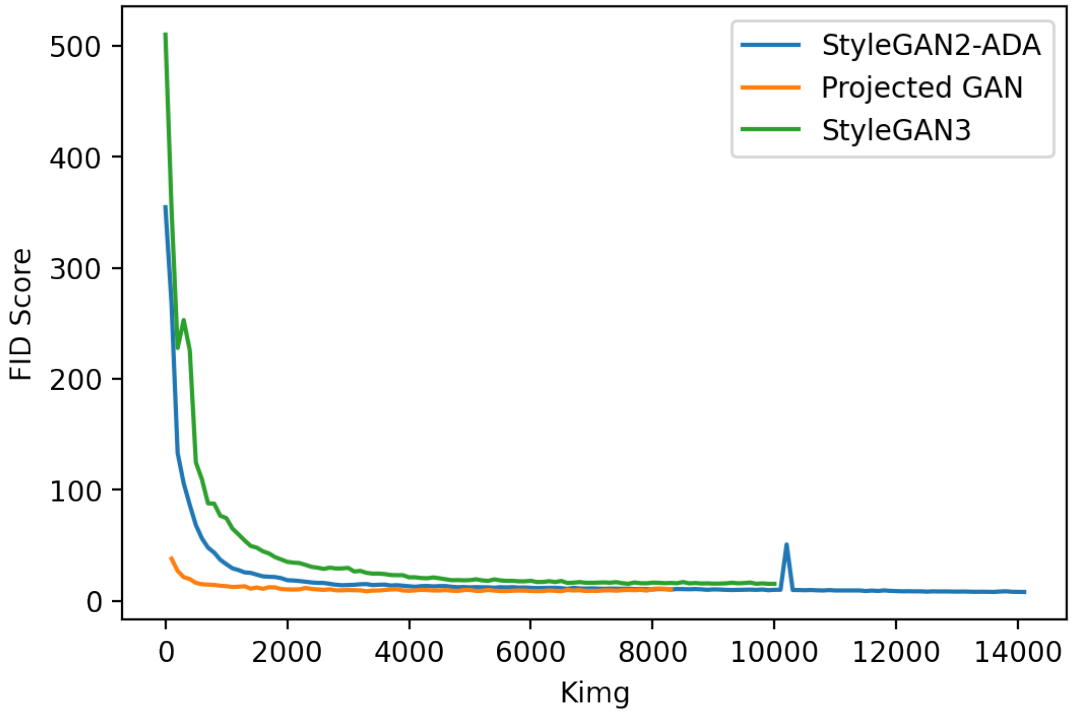


Figure 4.1: FID convergence for various GAN models (Binary-class).

Figure 4.1 depicts the learning curves of three GANs. The choice of the 'kimg' cut-off was made on the basis of the model convergence. As seen in the figure, in the initial stages ProjectedGAN performs well demonstrating faster convergence compared to StyleGANs. However, as the training progresses, StyleGAN performance surpasses ProjectedGAN. We, therefore, infer that for low-end GPU servers, ProjectedGAN is a favorable option as it requires less VRAM and computational power. The faster convergence of ProjectedGAN makes it an efficient choice during the early stages of training.

Figure 4.3, Figure 4.2, Figure 4.4 present illustrations of artworks generated by all each of the three models suggestive of positive and negative emotions.

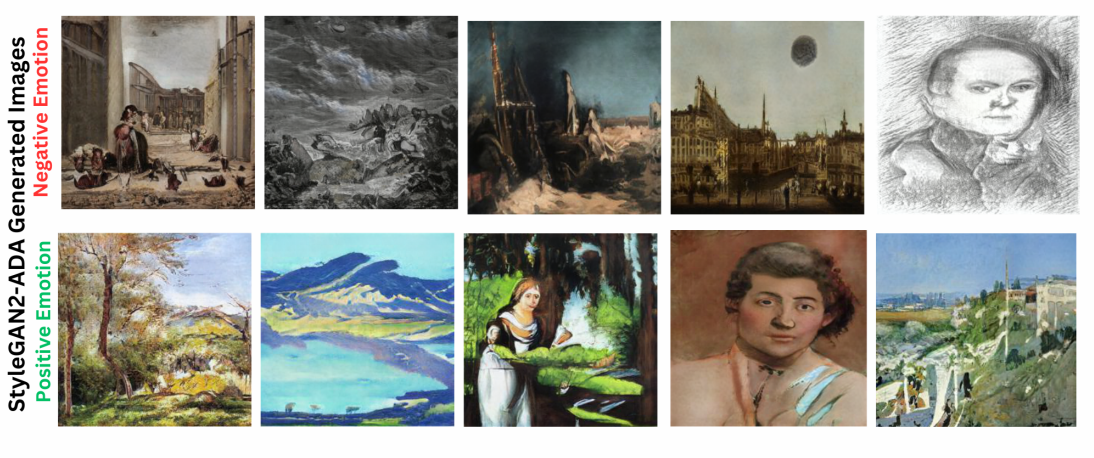


Figure 4.2: StyleGAN2-ADA: Generated artworks for the positive and negative emotion classes.

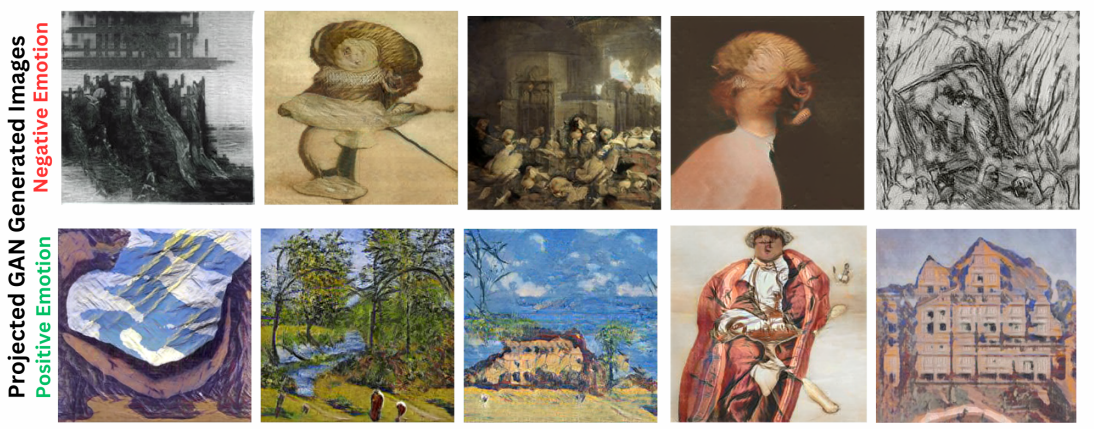


Figure 4.3: Projected GAN: Generated artworks for the positive and negative emotion classes.

Figure 4.4: StyleGAN3: Generated artworks for the positive and negative emotion classes.

4.1.2 Classification Models

We utilized Inception-ResNet-V2, ResNet50, VGG19, and Inception-V3 models for classification studies. 5-fold cross-validation was done to achieve reliable results knowing the stochastic nature of classification algorithms. By using different subsets of training and test data, the cross-validation strategy helps to mitigate stochasticity and achieve robust inferences.

The classification models were trained for a fixed number of epochs (30) with the Adam optimizer. We evaluated the accuracy, precision, recall, F1-score, MCC, ROC-AUC, and precision-recall curve for each fold. ROC curves summarize the trade-off between true and false positive rates across various probability thresholds, providing insights into the model's predictive performance. Similarly, precision-recall curves illustrate the exchange between the true positive rate and positive predictive value at different probability thresholds. A subset of 10,000 artworks was chosen from each emotion class based on the presence of a dominant emotion. Due to the balanced class distribution, the weighted and macro averages yielded similar results. For a balanced representation, 2,000 artworks were generated for each emotion class yielding 4,000 paintings per GAN.

Inception-ResNet-V2 outperformed all other classifiers on ROC-AUC, achieving a mean value 0.791 in five-fold cross-validation. Figure 4.5 presents the precision-curve and ROC for this Classifier. Table 4.2 provides an overview of the datasets, where the 'Test' column represents the test set obtained from the train-test split of the ArtEmis dataset. The Inception-ResNet-V2 model exhibits no notable class biases evident from its comparable performance on precision, recall, F1-score, and MCC score.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-score**	**MCC**
Test	0.716	0.716	0.716	0.716	0.432
Projected GAN	0.683	0.695	0.683	0.679	0.378
StyleGAN2-ADA	0.646	0.649	0.646	0.645	0.295
StyleGAN3	0.648	0.658	0.648	0.642	0.307

Table 4.2: Classification Performance of Inception-ResNet-V2: Accuracy, Precision, Recall, F1-score, and MCC for images generated by different models (Binary-class).

ResNet50-V2 demonstrated superior performance to other models on average accuracy, recall, precision, F1-score, and MCC score on the Test dataset reported in Table 4.3.

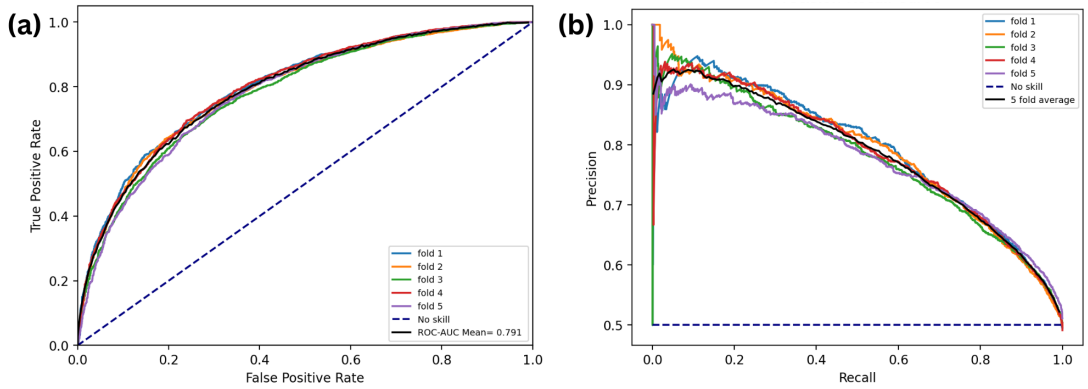


Figure 4.5: Classification Performance of Inception-ResNet-V2: (a) ROC Curve and (b) Precision-Recall Curve

The mean ROC-AUC for ResNet50-V2 was found to be 0.790, slightly lower than that of Inception-ResNet-V2. The details of ROC and Precision-recall curve are presented in Figure 4.6.

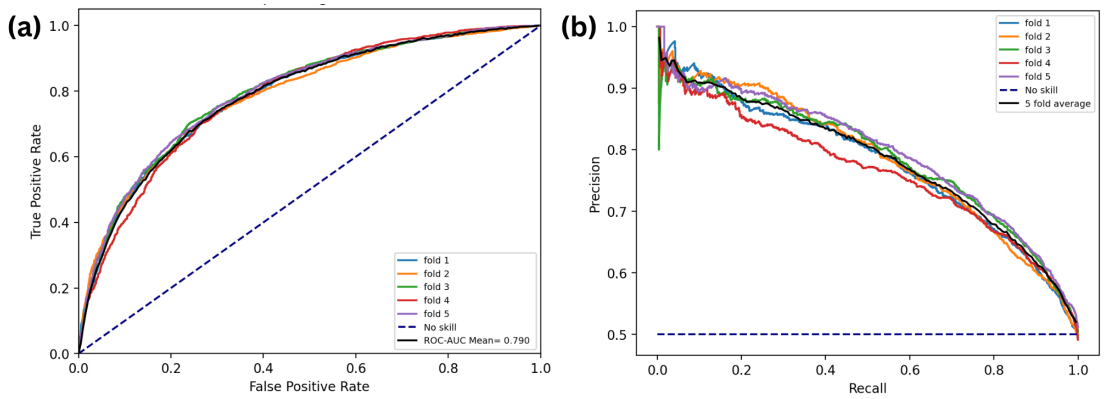


Figure 4.6: Classification Performance of ResNet50 V2: (a) ROC Curve and (b) Precision-Recall Curve.

Dataset	Accuracy	Precision	Recall	F1-score	MCC
Test	0.720	0.720	0.720	0.720	0.440
Projected GAN	0.699	0.703	0.699	0.698	0.402
StyleGAN2-ADA	0.658	0.658	0.658	0.658	0.317
StyleGAN3	0.652	0.661	0.652	0.648	0.313

Table 4.3: Classification Performance of ResNet50-V2: Accuracy, Precision, Recall, F1-score, and MCC for images generated by different models (Binary-class)

The average ROC-AUC for VGG19 was 0.768, indicating its superior performance in capturing the trade-off between true positive rate and false positive rate. Furthermore, VGG19 achieved an average accuracy of 69.9%. Detailed results for all evaluation metrics can be found in Table 4.4. The ROC and precision-recall curves for VGG19 are presented in Figure 4.7, providing additional insights into its classification performance.

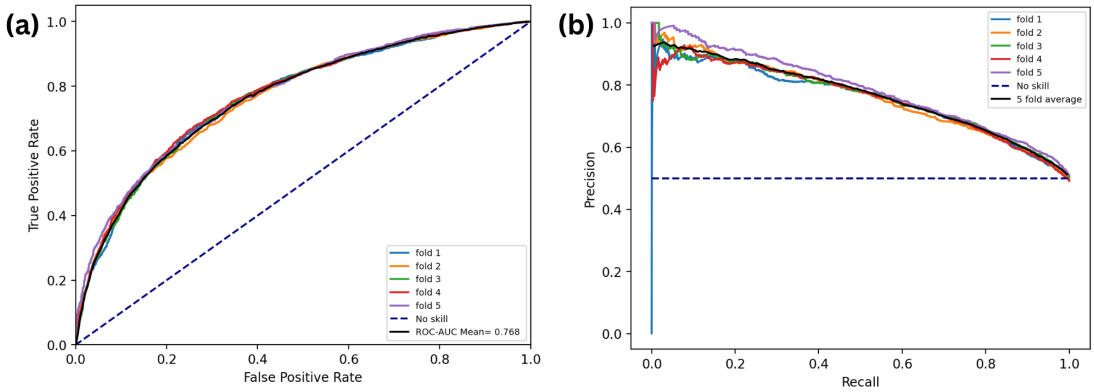


Figure 4.7: Classification Performance of VGG19: (a) ROC Curve and (b) Precision-Recall Curve.

Dataset	Accuracy	Precision	Recall	F1-score	MCC
Test	0.699	0.701	0.699	0.698	0.400
Projected GAN	0.697	0.706	0.697	0.693	0.402
StyleGAN2-ADA	0.654	0.656	0.654	0.653	0.311
StyleGAN3	0.652	0.658	0.652	0.693	0.309

Table 4.4: Classification Performance of VGG19: Accuracy, Precision, Recall, F1-score, and MCC for images generated by different models (Binary-class).

The Inception-v3 model achieved a mean ROC-AUC of 0.762 and an accuracy of 0.700. Detailed results for all evaluation metrics can be found in Table 4.5. Figure 4.5 showcases the ROC-AUC and Precision-recall curve specifically for the Inception-v3 model.

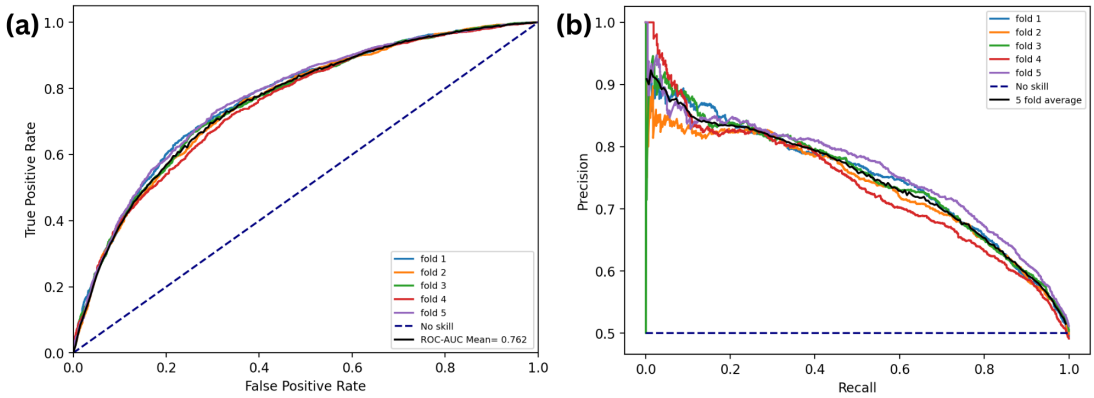


Figure 4.8: Classification Performance of Inception v3: (a) ROC Curve and (b) Precision-Recall Curve (Binary-class).

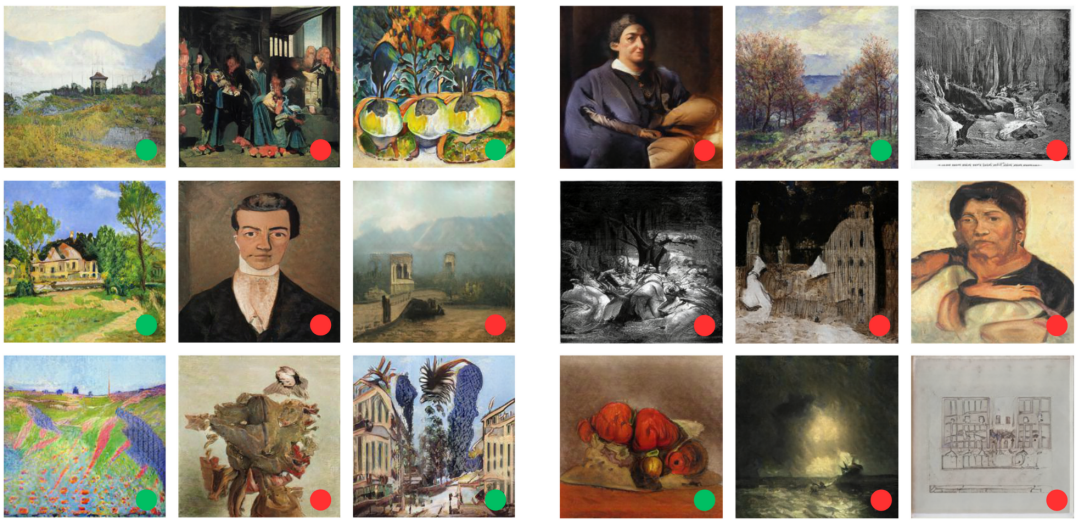
Dataset	Accuracy	Precision	Recall	F1-score	MCC
Test	0.700	0.700	0.700	0.700	0.400
Projected GAN	0.670	0.678	0.670	0.666	0.348
StyleGAN2-ADA	0.638	0.638	0.638	0.637	0.277
StyleGAN3	0.635	0.640	0.635	0.631	0.274

Table 4.5: Classification Performance of Inception-V3: Accuracy, Precision, Recall, F1-score, and MCC for images generated by different models (Binary-class).

As shown in Table 4.6, the ResNet50-V2 model outperformed other models achieving the highest accuracy of 0.720. Following closely behind, the Inception-ResNet-V2 model had an accuracy of 0.716, whereas the Inception-V3 and VGG19 models achieved an accuracy of 0.700 and 0.699, respectively. On the ROC-AUC metric, the Inception-ResNet-V2 model exhibited the best performance and achieved a score of 0.792. The ResNet50-V2 model followed closely with a score of 0.790, while the VGG19 and Inception-v3 models reached 0.768 and 0.762, respectively.

Classification Model	**Test**	**Projeted GAN**	**StyleGAN2-ADA**	**StyleGAN3**
Inception-ResNet-V2	0.716	0.683	0.646	0.648
ResNet50-V2	**0.720**	**0.699**	**0.654**	**0.652**
Inception-V3	0.700	0.670	0.638	0.635
VGG19	0.699	0.697	0.654	0.652

Table 4.6: Accuracy comparison of classification models corresponding to data generated three GAN model architectures (Binary-class).



(a) Generated images by StyleGAN2-ADA with Positive emotion **(b)** Generated images by StyleGAN2-ADA with Negative emotion

Figure 4.9: Illustrations of images of positive (a) and negative (b) classes generated by StyleGAN2-ADA. Green and red dots represent the Inception-ResNet-V2 classifier prediction as *Positive* and *Negative* emotion class, respectively.

Figure 4.9 showcases illustrations of images of positive and negative classes generated by StyleGAN2-ADA model and predictions made by Inception-ResNet-V2. We surmise that the classifier predicts the intended emotion with reasonably good accuracy for most of the generated artwork; however, there are occasional discrepancies in approximately 30-35% of the cases. This discrepancy suggests that the generation and classification models successfully capture relevant features of artworks. Going further, we conducted a human evaluation to assess the results from both the generative and classification models.

Aslan et. al. [32] analyzed the recognition of emotions evoked by artworks using visual features and knowledge graph embeddings and reported an accuracy of 81.57% for the binary classification. Notably, the logic of dominant emotion labeling used in this was distinctly different, and each artwork was assigned the most frequently occurring emotion. For instance, if an artwork receives annotations from 10 users, 6 indicating a *positive* emotion, 3 with a *negative* emotion, and 1 tag suggesting 'something else', it was labeled as having a *positive* emotion. Consequently, the dataset exhibits a significant class imbalance, with approximately 78% *positive* emotion artworks, 21% *negative* emotion artworks, and the remaining artworks falling into other categories. This class imbalance is also evident in the case of multiple classes. It is worth mentioning that this research article lacked inclusions of crucial evaluation metrics such as precision, recall, and F1-score essential for identifying potential biases within the classifier.

Considering evaluation metrics beyond accuracy is essential to address the class imbalance and potential biases. Reliance on 'accuracy' as a sole metric can lead to misleading interpretations. Our study utilized the ArtEmis V2.0 dataset, demonstrating less bias than ArtEmis V1.0. We employed 10,000 images per class and split the dataset into an 80:20 training-to-test ratio. Our accuracy on the test data was 72%, which is lower than the research above due to differences in the test dataset. However, we conducted a comprehensive evaluation using precision, recall, F1-score, ROC, and precision-recall curves to assess the model's performance.

Our model demonstrated resilience to class imbalance since we had an equal number of samples per class during training. The evaluation metrics, including accuracy, precision, recall, and F1-score, were similar, indicating a lack of class imbalance. Table 4.6 highlights that the ResNet50-V2 model outperforms other models, aligning with human evaluation discussed in Chapter 5. While the Inception-ResNet-V2 model achieved the best ROC-AUC score, there is a close competition between Inception-ResNet-V2 and ResNet50-V2.

4.2 Experiments with Multi-Class Paradigm

Following exploring the binary class problem, we extended our experiments to address the same problem in a multi-class setting. The multi-class problem involved a total

of five emotions. The data preprocessing and image selection steps were carried out as described in Chapter 3. The framework used for the multi-class problem remained consistent with the basic framework depicted in Figure 3.1. We trained three GANs: ProjectedGAN with FastGAN-lite generator, StyleGAN2-ADA, and StyleGAN3.

4.2.1 Generative Models

Generative Model	FID Score ↓	KID Score ↓	Precision	Recall
Projected GAN	**9.36**	**0.00136**	0.630	0.190
StyleGAN2-ADA	9.52	0.00233	**0.648**	0.240
StyleGAN3	10.41	0.00345	0.584	**0.359**

Table 4.7: Performance comparison of GAN models for multi-emotion class experiment.

Figure 4.10: StyleGAN2-ADA: Generated artwork generated for different emotion classes - (*amusement*, *awe*, *contentment*, *fear*, and *sadness*)

Table 4.7 presents the results obtained with various GAN models. The performance of models was evaluated using FID, KID, Precision, and Recall scores. The FID scores for ProjectedGAN, StyleGAN2-ADA, and StyleGAN3 were 9.36, 9.52, and 10.41, respectively. On KID metric, ProjectedGAN achieved the best performance with a score of 0.00136, followed by StyleGAN2-ADA with 0.00233, and StyleGAN3 with 0.00345. The ranking GANs based on FID and KID scores are consistent. It was observed that the model with a lower FID score also had a lower KID score, indicating better per-

Figure 4.11: Projected GAN: Generated artwork generated for different emotion classes - (*amusement*, *awe*, *contentment*, *fear*, and *sadness*)

Figure 4.12: StyleGAN3: Generated artwork generated for different emotion classes - (*amusement*, *awe*, *contentment*, *fear*, and *sadness*)

formance. The precision values for ProjectedGAN, StyleGAN2-ADA, and StyleGAN3 were 0.630, 0.648, and 0.538, respectively, suggesting that the generator produced high-quality images similar to real ones. The recall values for ProjectedGAN, StyleGAN2-ADA, and StyleGAN3 were 0.190, 0.240, and 0.359, respectively. A higher recall score indicates that the generator captured more features from the real data, resulting

in a diverse set of generated images. Initially, the models exhibited higher precision and lower recall, but as the number of epochs increased, they traded precision for recall, recognizing the importance of both aspects for effective image generation. The same computational resources were used for binary and multi-class classification experiments.

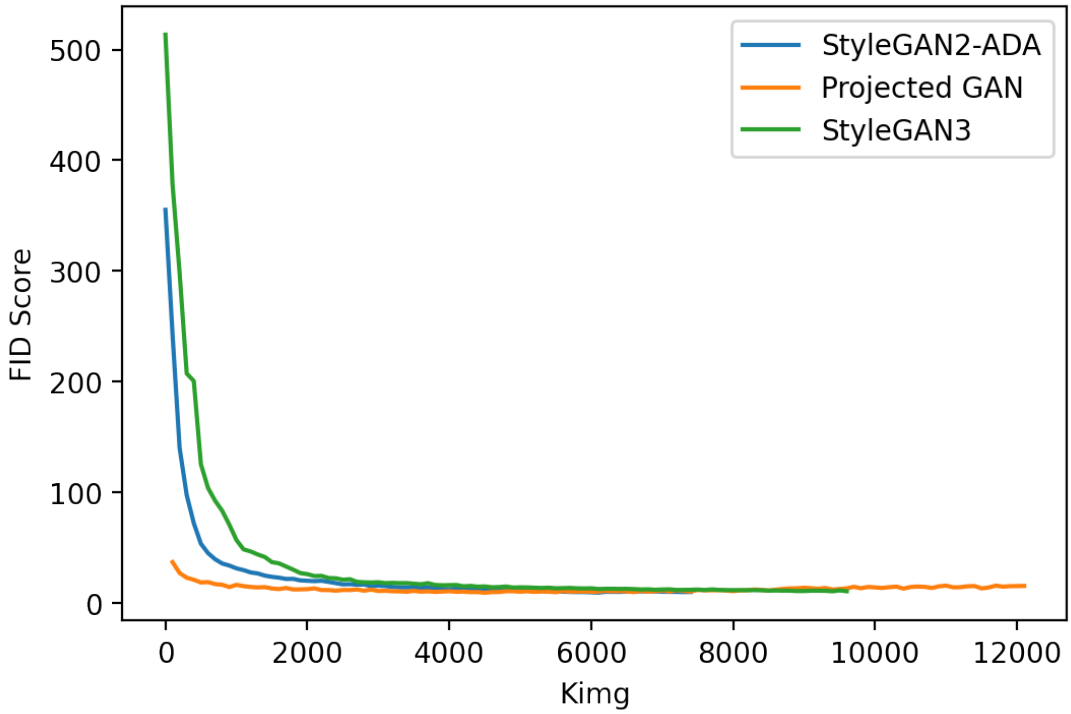


Figure 4.13: FID convergence for various GAN models (Multi-class)

Figure 4.13 shows the learning curves of three GANs pointing to the convergence models. The number of 'kimgs' was selected based on the convergence pattern, similar to the previous analogous experiment. In the case of multi-conditioned training, ProjectedGAN exhibited good performance and demonstrates faster convergence compared to StyleGANs. Therefore, for low-end GPU servers, Projected GAN should be a preferred option as it requires less VRAM and computational power. The efficient convergence of ProjectedGAN makes it well-suited for the initial stages of training. Overall, the learning curves of the three GANs overlap, indicating minimal differences between their performances.

Figure 4.10, Figure 4.11, and Figure 4.12 represent samples of generated artworks evoking specific emotional attributes.

4.2.2 Classification Models

In this study, we utilized widely recognized classification models (Inception-ResNet-V2, ResNet50, VGG19, and Inception-V3), similar to the binary paradigm. The classification models were trained for a fixed number of epochs (50) using 5-fold cross-validation with the Adam optimizer. Notably, we observed that models in multi-class scenarios required a higher number of epochs to reach the saturation point. The evaluation metrics including accuracy, precision, recall, and F1-score were assessed for each fold, and the average results across all five folds are presented in respective sections.

Our experiment focused on five specific emotion classes, aligning them with the GANs used: *Amusement*, *Awe*, *Contentment*, *Fear*, and *Sadness*. Initially, the models were evaluated using the test data, and subsequently, the same models were employed to assess the generated images from the GANs. Precision, recall, and F1-score were computed, and the weighted average was reported. We ensured a balanced representation of the generated images by generating 2,000 artworks for each emotion class, resulting in a dataset with 10,000 paintings per GAN.

Among all the classifiers, Inception-ResNet-V2 achieved the highest accuracy, precision, recall, and F1-score on the test dataset, with values of 54.6%, 52.7%, 54.6%, and 53.2% respectively, provided in Table 4.8.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-score**
Test	0.546	0.527	0.546	0.532
Projected GAN	0.480	0.544	0.480	0.455
StyleGAN2-ADA	0.389	0.454	0.389	0.346
StyleGAN3	0.369	0.423	0.369	0.336

Table 4.8: Classification Performance of Inception-ResNet-V2: Accuracy, Precision, Recall, and F1-score, for images generated by different models (Multi-class).

The test accuracy, precision, recall, and F1-score for ResNet50-V2 are reported as 53.4%, 52.2%, 53.7%, and 52.9%, respectively. The corresponding results of ResNet50-V2 can be found in Table 4.9.

The test accuracy, precision, recall, and F1-score for VGG-19 are reported as 52.7%, 50.6%, 52.7%, and 50.9%, respectively. The corresponding results of VGG-19 can be found in Table 4.10.

Inception-V3 achieved a test accuracy of 51.4%, precision of 50.2%, recall of 51.4%,

Dataset	Accuracy	Precision	Recall	F1-score
Test	0.537	0.526	0.537	0.529
Projected GAN	0.476	0.526	0.476	0.452
StyleGAN2-ADA	0.391	0.439	0.391	0.361
StyleGAN3	0.371	0.410	0.372	0.342

Table 4.9: Classification Performance of ResNet50-V2: Accuracy, Precision, Recall, and F1-score, for images generated by different models (Multi-class).

Dataset	Accuracy	Precision	Recall	F1-score
Test	0.527	0.506	0.527	0.509
Projected GAN	0.500	0.569	0.569	0.481
StyleGAN2-ADA	0.384	0.4	0.384	0.354
StyleGAN3	0.373	0.437	0.373	0.339

Table 4.10: Classification Performance of VGG19: Accuracy, Precision, Recall, and F1-score, for images generated by different models (Multi-class).

and F1-score of 50.6%. These results are presented in Table 4.11.

Dataset	Accuracy	Precision	Recall	F1-score
Test	0.514	0.502	0.514	0.506
Projected GAN	0.460	0.495	0.460	0.442
StyleGAN2-ADA	0.373	0.418	0.373	0.348
StyleGAN3	0.344	0.376	0.344	0.320

Table 4.11: Classification Performance of Inception-V3: Accuracy, Precision, Recall, and F1-score, for images generated by different models (Multi-class).

Despite the imbalance in the training data, no substantial difference was observed between the accuracy and weighted average of F1-scores. The 10,000 generated images from GANs (Projected GAN, StyleGAN2-ADA, StyleGAN3) were uniformly distributed across all classes. This equitable distribution results in smaller differences between accuracy, precision, recall, and F1-score. In the research, Aslan et. al. [32] implemented multi-emotion classification using 8 classes and achieved an accuracy of 45.39%. In the classification problem, it is well-known that as the number of classes increases the number of prediction options and the chances of making mistakes also increase. Consequently, increasing the number of classes in the classifier decreases accuracy, precision, recall, and the F1-score. We provide a range of evaluation metrics that suggest the absence of class bias in our model.

On the test dataset, Inception-ResNet-V2 outperformed all other classifiers as mentioned in Table 4.12. The table offers a comprehensive comparison of all classifiers

Classification Model	Test	Projected GAN	StyleGAN2-ADA	StyleGAN3
Inception-ResNet-V2	**0.546**	0.480	0.389	0.369
ResNet50-V2	0.537	0.476	**0.391**	0.371
Inception-V3	0.514	0.460	0.373	0.344
VGG19	0.527	**0.500**	0.384	**0.373**

Table 4.12: Accuracy comparison of classification models corresponding to data generated three GAN model architectures (Multi-class).

suggesting comparable performances. However, when examining the classification performance of the generated images, VGG19 demonstrated good performance in classifying Projected GAN and StyleGAN3 generated images, whereas ResNet50-V2 excelled in classifying StyleGAN2-ADA generated images. Furthermore, we observed that classifiers results were more accurate for the Projected GAN generated in both these paradigms.

4.3 Observations

According to the study by Brock et al. (2018) [42] and Heusel et al. (2017) [18], having more data can improve the Fréchet Inception Distance (FID) score of generative image models, as a more extensive dataset provides the model with more examples to learn from and can help to reduce overfitting. Similarly, Karras et al. (2017) [43] found that a larger dataset size can help to improve the quality of generated images and reduce the FID score.

Consistently, we find that increasing the number of training images per class can reduce the FID score, indicating that the generated images correlate more with the training images. It is consistent with the findings of Heusel et al. (2017) and Karras et al. (2017) mentioned above. However, as noted by Zhang et al. (2018) [44], many factors can affect the relationship between dataset size and FID score. It is important to consider these factors in optimizing the performance of generative image models.

Among the three GANs evaluated, StyleGAN2-ADA demonstrated a balanced performance in terms of precision and recall. It achieved good FID and KID scores, indicating high similarity between generated and real images and high precision and recall scores for both experiments. StyleGAN3 excelled in recall, surpassing the other models. A higher recall score implies that the model captures more features than the others, indi-

cating greater diversity in the generated images. Projected GAN performs well in FID, KID, and precision, but its recall score is significantly lower. It is a good model choice for scenarios with limited computational power.

In the context of GANs, precision, and recall are crucial metrics for evaluating performance. Precision assesses the accuracy of the generator in producing realistic images that closely resemble the real ones. A higher precision signifies high-quality image generation. On the other hand, recall measures the proportion of high-quality samples in the generated dataset, reflecting the facility of GANs to capture important features from the original dataset. Precision and recall present a trade-off between image quality and diversity. A higher precision may come at the expense of diversity, while a higher recall may sacrifice image quality. Striking a balance between precision and recall is essential for generating high-quality and diverse images [21]. We examined the trade-offs of our three GANs. Projected GANs prioritize quality over diversity, whereas StyleGAN3 prioritizes diversity over quality. On the other hand, StyleGAN2-ADA strikes a favorable trade-off by achieving a balance between precision and recall.

Furthermore, a correlation is observed between the precision of a GAN and the accuracy achieved by classifiers for the artworks generated by that GAN. Higher precision results in higher accuracy for the generated images. This correlation can be attributed to the fact that the classifier evaluates whether an image falls within the probability distribution of the target class.

Among the classifiers evaluated, ResNet50-V2 emerges as the top performer, surpassing the other classifiers when considering multiple metrics collectively. In addition, a human evaluation was conducted to assess the quality of the GANs and classifiers, as detailed in Chapter 5.

CHAPTER 5

Human Evaluations: Turing Test and Guess the Emotion

Human validation is crucial in image generation problems to assess the quality, realism, and perceptual aspects of generated images. While quantitative metrics such as FID and KID provide objective evaluations, they do not fully capture the nuances of human perception. Incorporating human validation allows for a more comprehensive assessment of the generated images.

A study by Salimans et al. (2018) [44] emphasized the significance of human evaluation in image generation research. They compared GAN models using quantitative metrics and human ratings, finding that human judgments captured perceived image quality and diversity more accurately than automated metrics. The need for Turing testing arises from the desire to assess and evaluate the capabilities of artificial intelligence (AI) systems, particularly in human-like intelligence and behavior. The Turing test, named after British mathematician and computer scientist Alan Turing, is designed to determine if a machine can exhibit intelligent behavior indistinguishable from a human [45]. It is crucial in assessing and improving the ethical considerations surrounding AI systems, helping us understand their capabilities, limitations, and potential societal implications.

We have developed an experimental setup called the 'Turing Test for Artist' (TTA) to assess the authenticity and emotional impact of generated art images. We implemented this test only for binary classification – fake (computer generated) vs. real paintings and *positive* vs. *negative* emotion. We utilized GAN models to generate images as discussed in the preceding sections. To evaluate the quality of our generator, we have implemented the TTA, which involves human evaluation of artistic creations. Participants are instructed to rate the authenticity of each image on a scale of 0 to 5, where a rating of 0 indicates that the image appears fake or generated. In contrast, a rating of 5 implies that it seems real. Additionally, participants are asked to assess the dominant emotion evoked by the art on a scale of 0 to 3, where a rating of 0 represents a *negative* emotion, and a rating of 3 represents a *positive* emotion. The TTA and the Guess the Emotion tasks can be accessed at the following link: `https://cosylab.iiitd.edu.in/tta/`)

5.1 Design of Experiments

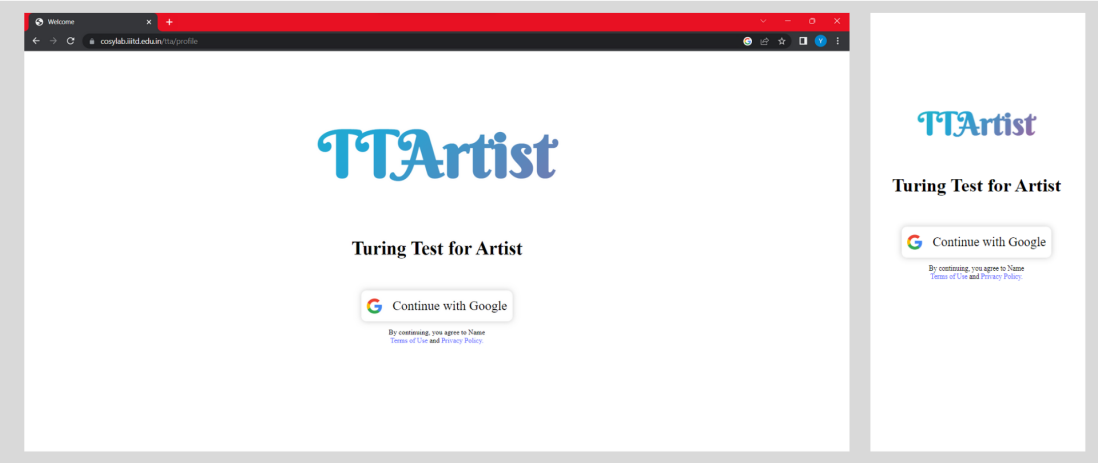


Figure 5.1: Turing Test for Artist: 'Login/Signup Page View' for both desktop and mobile versions.

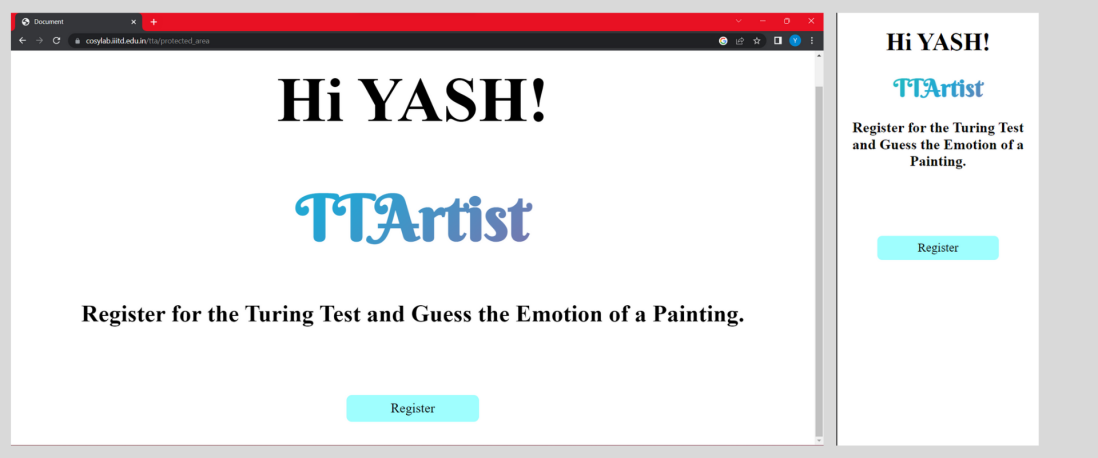


Figure 5.2: Turing Test for Artist: 'Registration Page view' for desktop and mobile versions.

From the user's perspective, this research experiment requires participants to enroll by providing their email addresses, as shown in Figure 5.1.

Subsequently, the participants receive instructions about the experiment as shown in Figure 5.3. Upon enrollment, the test commences by offering users two options; to submit their response or to skip the image as shown in Figure 5.4. If the user wishes to provide annotations for both options, they can make their selection and click the submit button. However, if the user finds an image confusing and cannot judge, they can skip it. Participants can access their statistics in the profile section, which includes the total number of evaluated images and the number of images skipped. The statistics

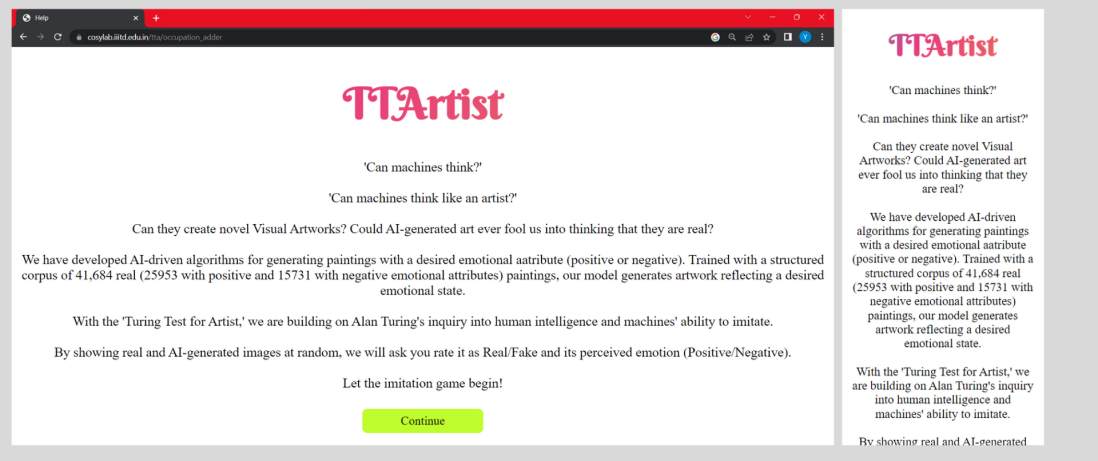


Figure 5.3: Turing Test for Artist: 'Instructions Page View' for both desktop and mobile versions.

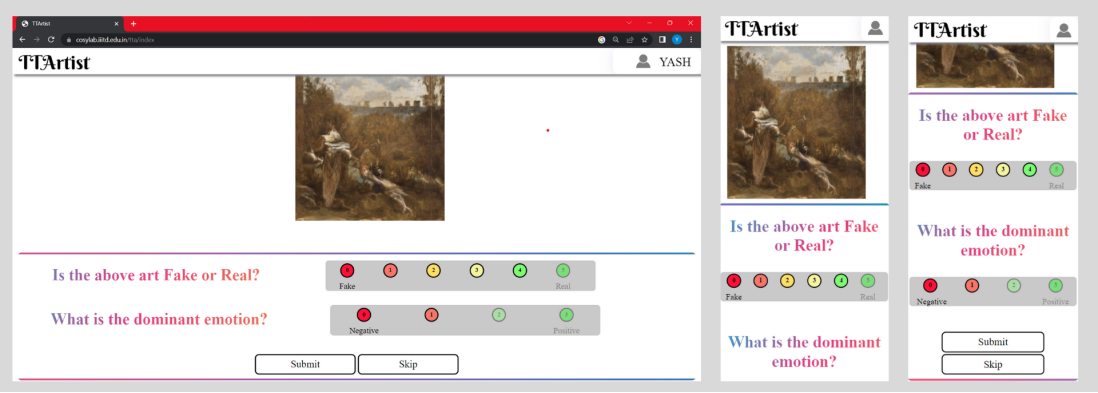


Figure 5.4: Turing Test for Artist: 'Main Page View' for both desktop and mobile versions.

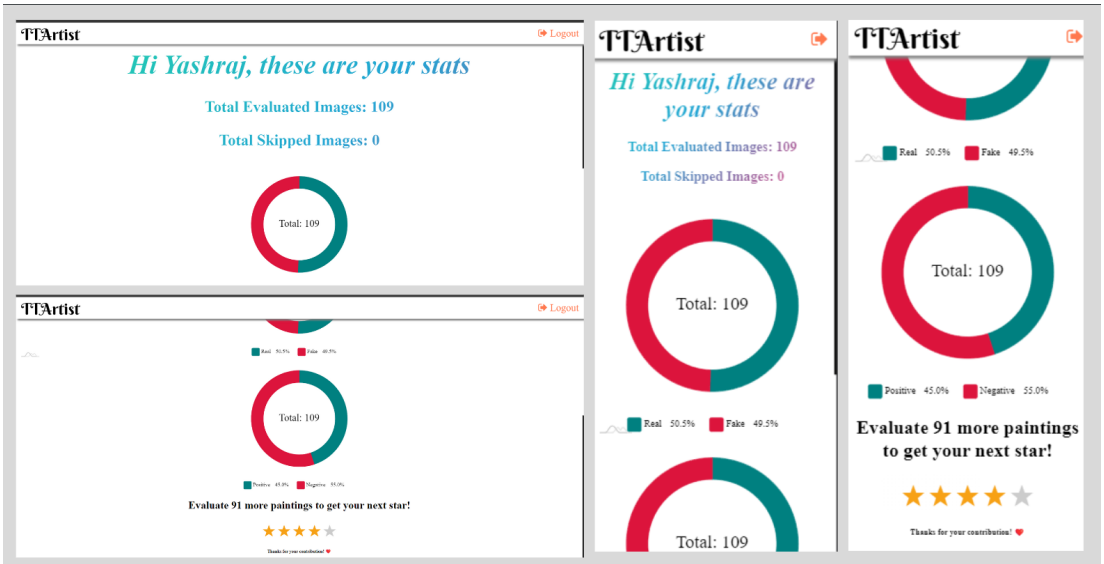


Figure 5.5: Turing Test for Artist: 'User Statistics Page View' for both desktop and mobile versions.

section also provides information about the proportion of real and generated images and that of *positive* and *negative* images presented to the user as shown in Figure 5.5. To incentivize participation, users are awarded ratings based on the number of pictures they evaluate. The more the number of pictures evaluated, the more the number of stars the user is awarded.

5.1.1 Data Sampling for Turing Test

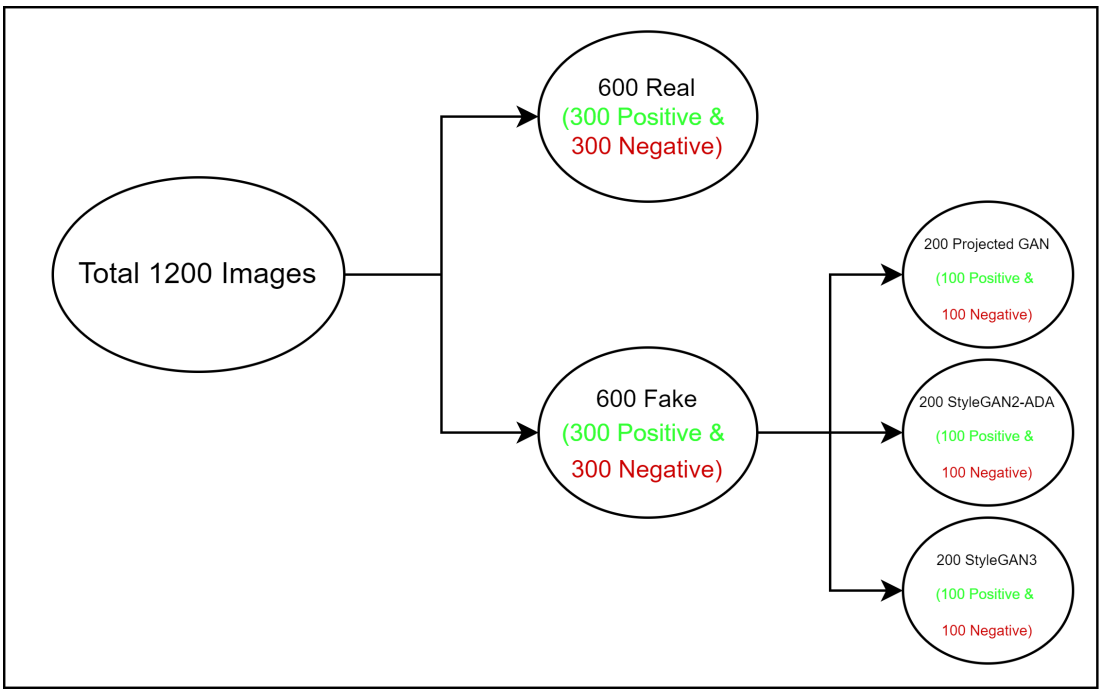


Figure 5.6: Turing Test for Artist: Distribution of dataset class.

For TTA dataset, we have considered a total of 1200 art images. All the generated images belong to the binary-class experiment. Each image has two attributes: generated or real and *positive* or *negative* emotion. Among the 1200 images, 600 are real, and 600 are fake. Among the 600 fake images, 200 belong to Projected GAN, 200 to Stylegan3, and 200 to Stylegan2-ADA. Therefore, when considering the distribution of emotions, out of the total 1200 images, 600 pertain to the *positive* class and 600 to the *negative* class. Within the *positive* emotion class, 300 images are real, while 100 images belong to the three GANs. For generated image, if an image is generated to have *positive* emotion, then we are considering this image as a *positive* emotion class image. Similarly, within the *negative* emotion class, 300 images are real, while 100 images each belong to the three GANs. A detailed flowchart of dataset distribution is explained in Figure 5.6

where Real/Fake represents whether the image is fake or real. *Positive* and *negative* represents emotion class. All image selections are made randomly and impartially.

5.1.2 Tech Stack

- Front-end Framework: HTML, CSS, Javascript (Vanilla), am4charts
- Back-end Database Management: Sqlite3
- Integration of Front-end and Back-end: Flask, python
- Data Retrieval and Render: Jinja and Ajax
- Authentication: Google Oauth API
- Server Hosting: Docker

5.1.3 SQLite Database Structure

The name of the database is 'data.db'. The database has 3 tables: user, image, and user_evaluation. The 'image' table contains information about the dataset. The 'user' table stores the information of the user. The 'user_evaluation' table contains the information of 1 evaluation provided by the user on a particular image. The schema of the databases are following:

CREATE TABLE Images (imgId int NOT NULL PRIMARY KEY, filepath varchar(256), fake_or_real varchar(256), emotion int);

- imgId: ImageId for an image. It is not a null primary key.
- filepath: Contains the path of the file as a string.
- fake_or_real: Contains whether the image is fake or real. If it is real, it is tagged as real. If the given image is generated, it is associated by its GAN name Projected GAN, stylegan2, and stylegan3.
- emotion: Contains 0 and 1 where 0 represents *negative* emotion and 1 represents *positive* emotion.

CREATE TABLE user_evaluation(sno INTEGER PRIMARY KEY AUTOINCREMENT, uid varchar(256), imgId int, fake_or_real int, emotion int);

- sno: Serial number of the evaluation.
- uid: Contains the user ID who evaluates the particular image.

- imgId: ID of the image which is evaluated.
- fake_or_real: Contains an integer between 0 to 5 where 0 represents generated/fake and 5 represents the real given by the user.
- emotion: Store integer between 0 to 3 where 0 represents *negative* and 3 represents the real given by the user.

CREATE TABLE user(uid varchar(256) NOT NULL PRIMARY KEY, email varchar(256), name varchar(256), random_counter int, total_count int,total_skip int, total_pos int, total_real int);

- Uid: Unique user ID is given by Google OAuth. The primary key of the table.
- Email: Email ID of the user.
- Name: Name of the user given by Google OAuth.
- random_counter: It is the random image id assigned to the user.
- total_count: Total number of images evaluated.
- total_skip: Total number of skipped images.
- total_pos: Total number of *positive* emotion classes shown to the user.
- total_real: Total number of real images shown to the user.

5.1.4 Data Preprocessing and Protocols

- **Binarization of Evaluator Assessment**: The Turing Test for Artist collects human assessment of an image for it being fake or real on a scale of 0 to 5. For the purpose of binary analysis, ratings between 0 to 2 were identified as fake, while those with ratings between 3 and 5 were labeled as real. Similarly, in the 'Guess the Emotion' task, we identify ratings 0 and 1 as *negative* emotions, while ratings 2 and 3 correspond to *positive* emotions.
- **Random Data Sampling**: In order to maintain the randomness and eliminate class biases in the experiment, several measures were implemented. All 1200 images were randomly selected and distributed equally across the classes.
- **Random Presentation of Data**: A shuffled list of images was entered into the 'image' table. When a new user signs up, a random number between 0 to 1199 is generated and assigned to the 'random_counter' variable. After each evaluation or image skip, the value of the random_counter is incremented by 1. The image with (random_counter + 1) % 1200 ID is presented to the user for evaluation, ensuring a diverse sequence of images.

- **Unique Data Presentation**: Measures were implemented to ensure a user cannot evaluate the same image twice.
- **Stopping Criterion for the Evaluator**: Once any user completes the evaluation for all 1200 images, the evaluation process for the user is declared complete.
- **Security**: From a security perspective, precautions were taken to prevent SQL injection attacks on the database.
- **Deployment**: In terms of hosting and deployment, Docker was employed to facilitate easy hosting and provide faster deployment and migration.

5.2 Results of the Turing Test (Fake or Real)

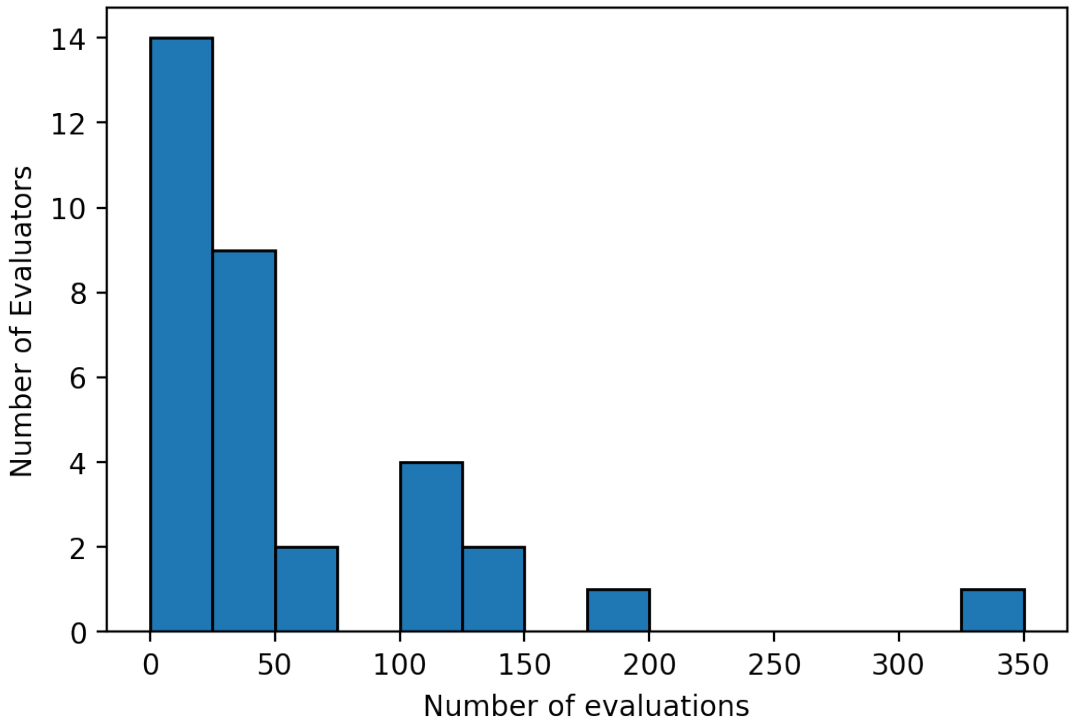


Figure 5.7: Turing Test for Artists: Distribution of annotations by the evaluators.

A total of 33 users conducted a total of 1708 evaluations as distribution shown in Figure 5.7. Most users were Indraprastha Institute of Information Technology, Delhi (IIITD) students. However, some users provided few ratings, prompting us to apply a cutoff for user selection. We considered users who had completed at least 25 evaluations as valid. We implemented this criterion to ensure that users evaluated a sufficient number of images and to gather meaningful insights regarding their commitment to the

test. Additionally, we examined whether users evaluated images randomly or not. Due to the limited number of evaluations for each user, it was not feasible to analyze individual users, leading us to apply the cutoff. To avoid bias, we excluded our submissions from the evaluations.

Assumptions:-

- If a fake image is tagged as fake: True Negative
- If a real image is tagged as fake: False Negative
- If a fake image is tagged as real: False Positive
- If a real image is tagged as real: True Positive

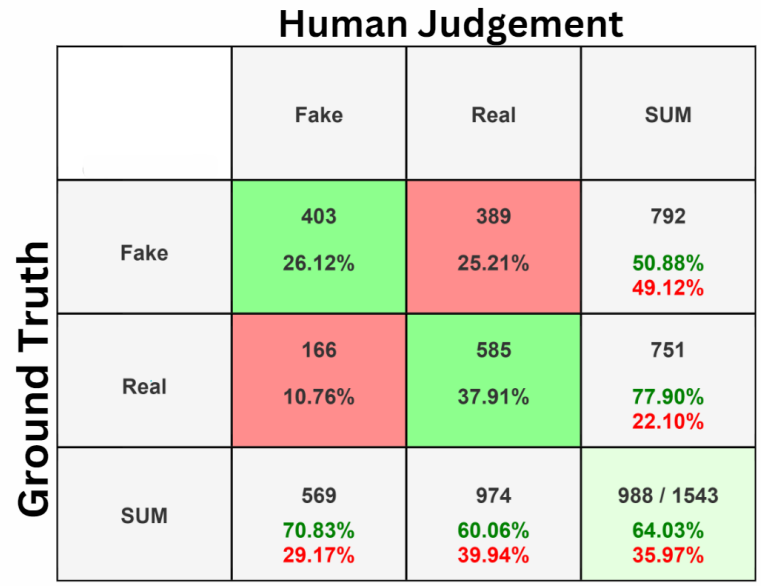


Figure 5.8: Confusion Matrix for the Turing Test for Artist.

The total number of valid evaluations, excluding our submissions, amounted to 1543 with 18 valid users. On average, a user evaluates 85.72 images. The confusion matrix illustrates the distribution of predictions and ground truth, with 403 True Negatives (TN), 389 False positives(FP), 585 True Positives(TP), and 166 False Negatives (FN). The False Positive Rate (FPR = FP/(FP+TN)) for the Turing test is 49.12%, and the accuracy is 64.03%. The True positive rate (TPR = TP/ (TP + FN)) for the Turing test is 77.90%.

Regarding the model-wise statistics:

- For Projected GAN, a total of 270 valid evaluations were conducted, with 134 evaluations labeling the images as real and 136 evaluations considered fake. This indicates that 49.63% of the generated images were labeled as real.

- For StyleGAN2-ADA, 267 valid evaluations were conducted, with 128 predictions of fake and 139 real predictions. Thus, 52.06% of the generated images by StyleGAN2-ADA were labeled as real.
- For StyleGAN3, a total of 255 valid evaluations were conducted, with 139 predictions of fake and 116 predictions of real. Consequently, 45.49% of the generated images by StyleGAN3 were labeled as real.

5.2.1 Observations

Users with high true positive (TP) and true negative (TN) counts are potentially experts, indicating their ability to distinguish between fake and real images. In our study, the TP count was 585, and the TN count was 403, significantly higher than the false negative (FN) count of 166. This suggests that the average user possessed sufficient knowledge to perform the experiment.

A high false positive (FP) count (403) indicates that our generative models successfully deceived users. With an FPR of 49.12%, the generator fooled users almost half the time. It is natural to question if users randomly labeled images as fake or real when the FPR is close to 50%. However, this can be explained by the high TPR of 77.90%. The TPR is significantly higher than 50%, demonstrating that users took the experiment seriously and did not randomly evaluate images.

Ideally, we aim to deceive users, resulting in high FP and low FN counts, while users should have high TP and TN counts to demonstrate their knowledge and ability to distinguish between real and fake images. In our case, we observed an FP count of 403 and an FN count of 166, indicating that the model deceived the average user. Additionally, the high TP and TN counts suggest that the average user was knowledgeable. Combining these observations, we can conclude that the model successfully fooled genuine users in many cases.

When considering individual model performance, StyleGAN2-ADA exhibited the best false positivity ratio. StyleGAN2-ADA deceived users 52.06% of the time, followed by Projected GAN at 49.63% and StyleGAN3 at 45.49%. This sequence aligns with the FID score rankings provided in Table 4.1, where StyleGAN2-ADA outperformed Projected GAN and StyleGAN3. Therefore, based on quantitative, qualitative, and human

validation results, StyleGAN2-ADA emerges as the superior model among the three.

5.3 Results for 'Guess the Emotion' Test

In our study, we concurrently collected evaluations to guess the dominant emotion. The number of evaluations and users remained the same, as mentioned in the previous section, and we applied the cutoff criterion of at least 25 evaluations per user. However, the task of guessing the dominant emotion poses inherent ambiguity. It can vary from person to person, gender to gender, and place to place [34]. In the context of ArtEmis, a single artwork can evoke multiple emotions, further emphasizing its ambiguous nature. Consequently, our users' perception of guessing the emotion might not align with the average annotator of ArtEmis.

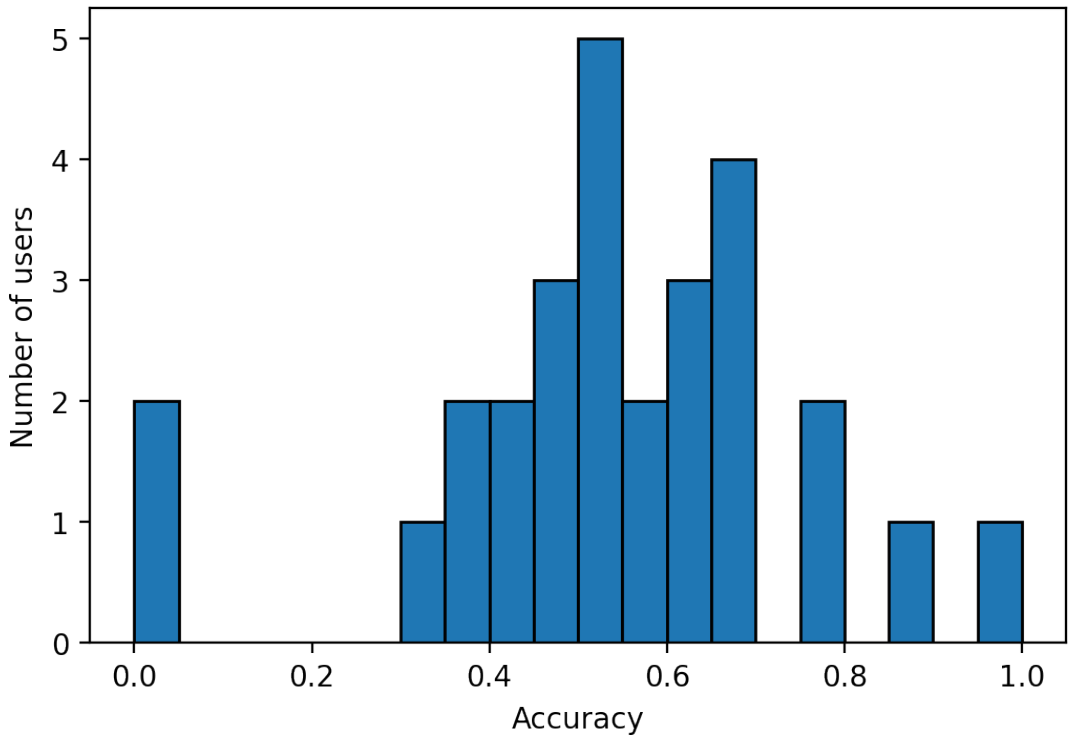


Figure 5.9: Turing Test for Artist: Distribution of accuracy corresponding to the real data and evaluators' assessment.

To address this issue, we introduced an additional cutoff criterion to select evaluators based on the accuracy of their emotion evaluations on real images. For getting selected, the accuracy of guessing the emotion on real data should be greater than or equal to a certain threshold. For instance, if a user evaluates 100 images, consisting of 50 real and 50 fake ones, we assess the accuracy of the evaluator's emotion tags compared to

the ArtEmis annotations (considered ArtEmis as ground truth). Suppose the evaluator correctly tags the emotions of 30 real images; then the accuracy would be calculated as 30/50, resulting in 0.60. Since all generative and classification models are trained on the ArtEmis dataset, ensuring that our annotators' thought processes align with those of an average ArtEmis annotator is crucial. Given the relatively limited domain of our evaluator population, divergent perspectives could lead us astray. Therefore, it becomes necessary to apply different selection cutoff criteria. For each model, we explored various cutoffs based on evaluations of real images and then calculated the corresponding statistics for the generated images. In Figure 5.9, it can be observed that most of the evaluators belong between 0.35 to 0.70 accuracy.

When calculating the accuracy of our evaluators, it is essential to establish ground truth. For real images, we consider the ArtEmis annotations as the ground truth. However, a question arises regarding the ground truth for generated images. For instance, if a GAN generates an image intended to have *positive* emotion, but a classifier suggests it as *negative*, which source should be relied upon? Which is more reliable in classifying the image, the classifier or the GAN? In this section, we address this issue by considering different sources as the ground truth, including all classifiers and the GAN, and calculate accuracies based on these varying ground truths.

5.3.1 Projected GAN

Cutoff	Total Evaluation	Inception-ResNet-V2	ResNet50-V2	VGG19	Inception-V3	GAN
0.0	270	0.5704	**0.5963**	0.5259	0.5556	0.5481
0.40	260	0.5808	**0.5962**	0.5308	0.5615	0.5615
0.50	176	0.5739	**0.6136**	0.5398	0.5568	0.5682
0.55	143	0.6154	**0.6434**	0.5734	0.6084	0.5874
0.60	124	0.6452	**0.6532**	0.5887	0.6048	0.6048
0.65	108	0.6667	**0.6852**	0.6389	0.6296	0.6204

Table 5.1: Accuracies for various cut-offs of different ground truths. Generated images: Projected GAN. Ground Truth: Various classification models and Projected GAN.

The results presented in Figure 5.10 and Table 5.1 clearly demonstrate that increasing the cutoff for real images improves the accuracy of the generated images. Among the classifiers, ResNet50-V2 exhibits the highest accuracy, indicating that the emotion portrayed in the projected GAN-generated images aligns most closely with human perceptual judgment when using ResNet50-V2, followed by Inception-ReNnet-V2. Fur-

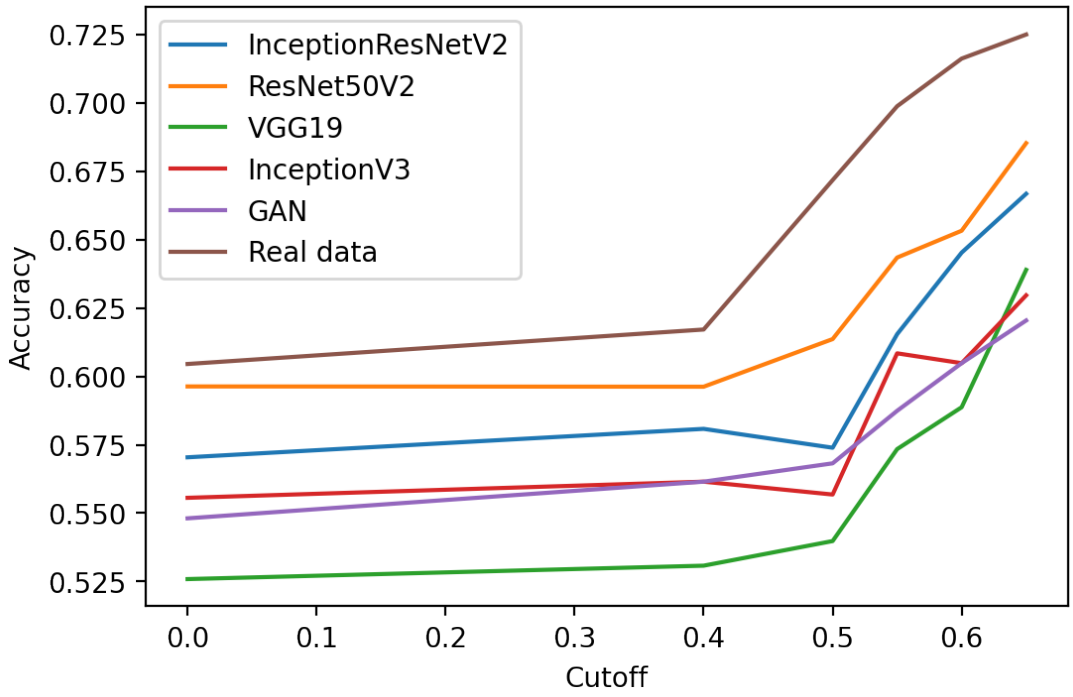


Figure 5.10: Projected GAN: Accuracies for various cut-offs of different ground truths.

thermore, VGG19 surpasses InceptionV2 at higher cutoffs. Notably, the number of evaluations decreases significantly with increased accuracy cutoff.

5.3.2 StyleGAN2-ADA

Cutoff	Total Evaluation	Inception-ResNet-V2	ResNet50-V2	VGG19	Inception-V3	GAN
0.0	267	0.5543	0.5543	**0.5693**	0.5280	0.5618
0.40	258	0.5543	0.5620	**0.5698**	0.5310	**0.5698**
0.50	190	0.5737	0.5895	0.5842	0.5737	**0.5947**
0.55	172	0.5930	**0.6105**	0.5930	0.5814	0.5988
0.60	150	0.5933	**0.6133**	0.600	0.5733	0.6067
0.65	139	0.5827	**0.6043**	0.5971	0.5683	**0.6043**

Table 5.2: Accuracies for various cut-offs of different ground truths. Generated images: StyleGAN2-ADA. Ground Truth: Various classification models and StyleGAN2-ADA.

Similar to projected GAN, the accuracy of StyleGAN2-ADA increases across all ground truth measures, as evident from the Table 5.2 and Figure 5.11. However, the accuracy differences among the four classifiers are not as pronounced as in projected GAN. Unlike the Projected GAN and StyleGAN3 model, all classifiers and GAN perform similarly for various cutoffs, suggesting no clear winner among them.

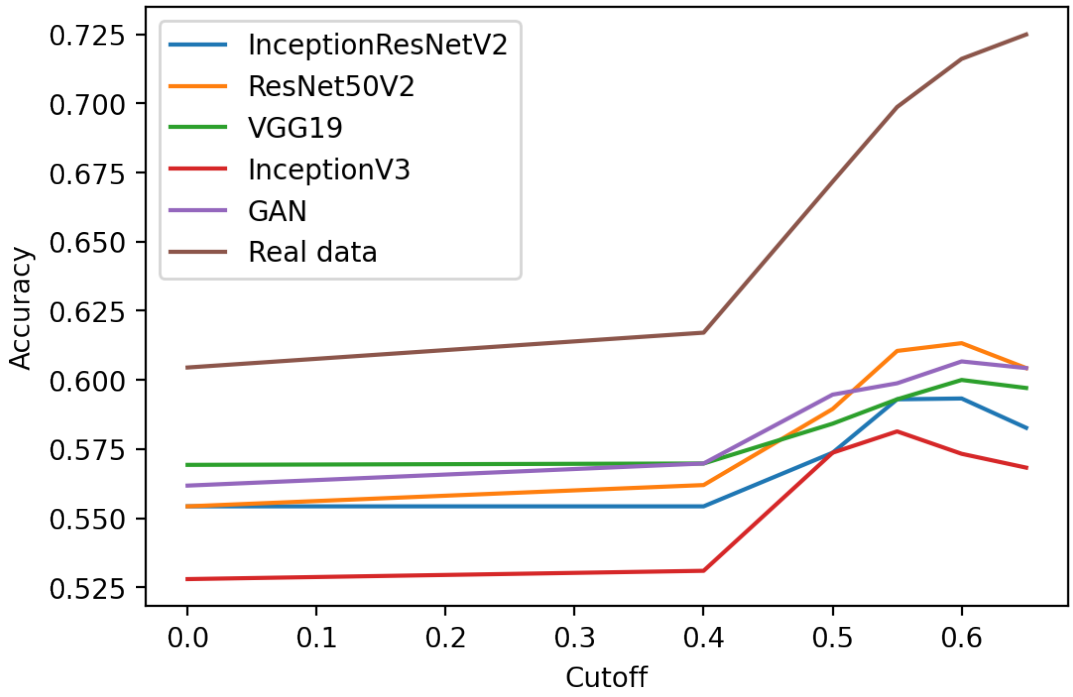


Figure 5.11: StyleGAN2-ADA: Accuracies for various cut-offs of different ground truths.

5.3.3 StyleGAN3

Cutoff	Total Evaluation	Inception-ResNet-V2	ResNet50-V2	VGG19	Inception-V3	GAN
0.0	255	0.5567	0.5529	0.5255	**0.5608**	0.4706
0.40	242	0.5496	**0.5620**	0.5331	0.5537	0.4711
0.50	174	0.5690	**0.6264**	0.5517	0.5862	0.4713
0.55	151	0.5762	**0.6556**	0.5497	0.5894	0.4635
0.60	131	0.5878	**0.6565**	0.5649	0.5878	0.4809
0.65	115	0.5913	**0.6522**	0.5652	0.5913	0.4957

Table 5.3: Accuracies for various cut-offs of different ground truths. Generated images: StyleGAN3. Ground Truth: Various classification models and StyleGAN3.

Table 5.12 and Figure 5.4 show the apparent relation between cutoff and accuracies like all other GANs. For lower cutoffs, ResNet50-V2, Inception-V3, and Inception-ResNet-V2 demonstrate comparable performance, but at higher cutoffs, ResNet50-V2 outperforms the other GAN models. Inception-V3 and Inception-ResNet-V2 exhibit similar performances. Notably, the disparity between classifier and GAN accuracies is significantly higher for StyleGAN3 than for the Projected GAN and StyleGAN2-ADA.

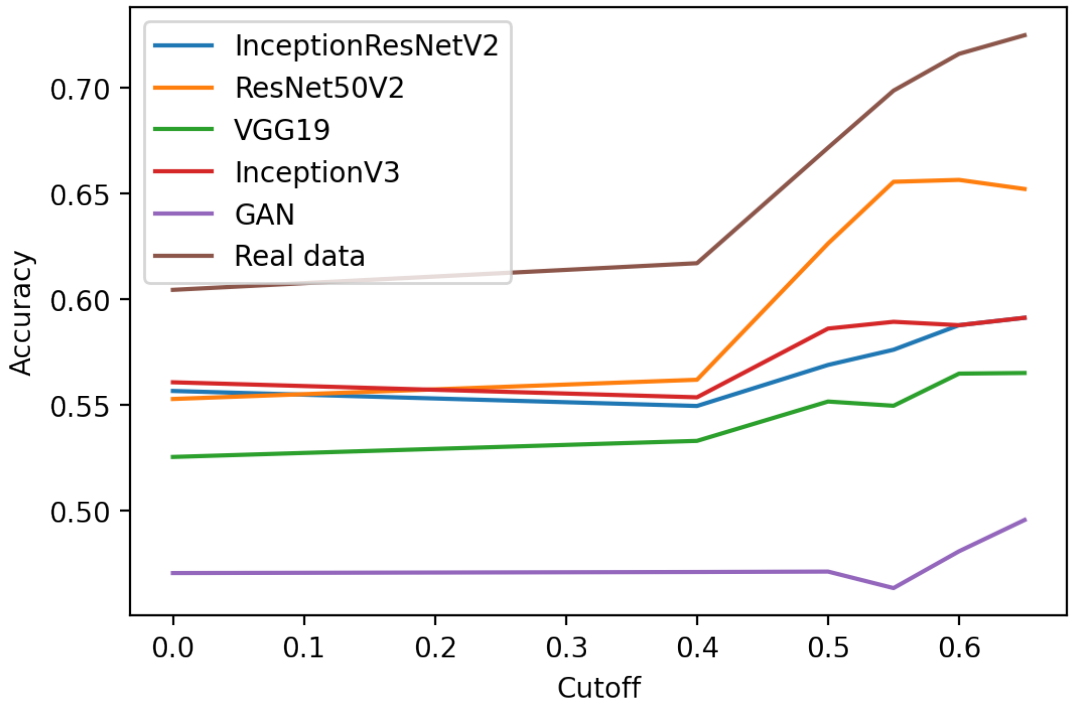


Figure 5.12: StyleGAN3: Accuracies for various cut-offs of different ground truths.

Cutoff	Total Evaluations	Total evaluators	Inception-ResNet-V2	ResNet50-V2	VGG19	Inception-V3	GAN	Real data
0.0	1543	18	0.5606	**0.5682**	0.5404	0.5480	0.5278	0.6045
0.40	1473	16	0.5618	**0.5737**	0.5447	0.5487	0.5355	0.6171
0.50	1064	12	0.5722	**0.6093**	0.5593	0.5722	0.5463	0.6717
0.55	914	10	0.5944	**0.6352**	0.5730	0.5923	0.5515	0.6987
0.60	789	8	0.6074	**0.6395**	0.5852	0.5877	0.5654	0.7161
0.65	711	6	0.6105	**0.6436**	0.5994	0.5939	0.5746	0.7249

Table 5.4: Combined results showing accuracies for various cut-offs of different ground truths. Generated images: From all GANs. Ground Truth: Various classification models and relevant GAN.
Combined results showing accuracies on varying ground truths

5.3.4 Observations

- In this section, we evaluated both real and generated images. In Table 5.4, real image evaluation is provided in the 'Real data' column, and it contains the accuracy of the user's evaluation with ArtEmis (Ground truth). For the classifiers (ResNet50-V2, Inception-V3, Inception-ResNet-V2, VGG19). 'GANs' are referring to the evaluations of all generated images. Therefore, the columns for all classifiers and GANs share the same set of evaluations, while the 'Real data' column specifically represents evaluations conducted on real images. The evaluations for all generated images, including all GANs, are summarized in Table 5.4 and visualized in Figure 5.13.

- Among the classification models, ResNet50-V2 achieves the highest accuracy, followed by Inception-ReNet-V2, Inception-V3, and VGG19.

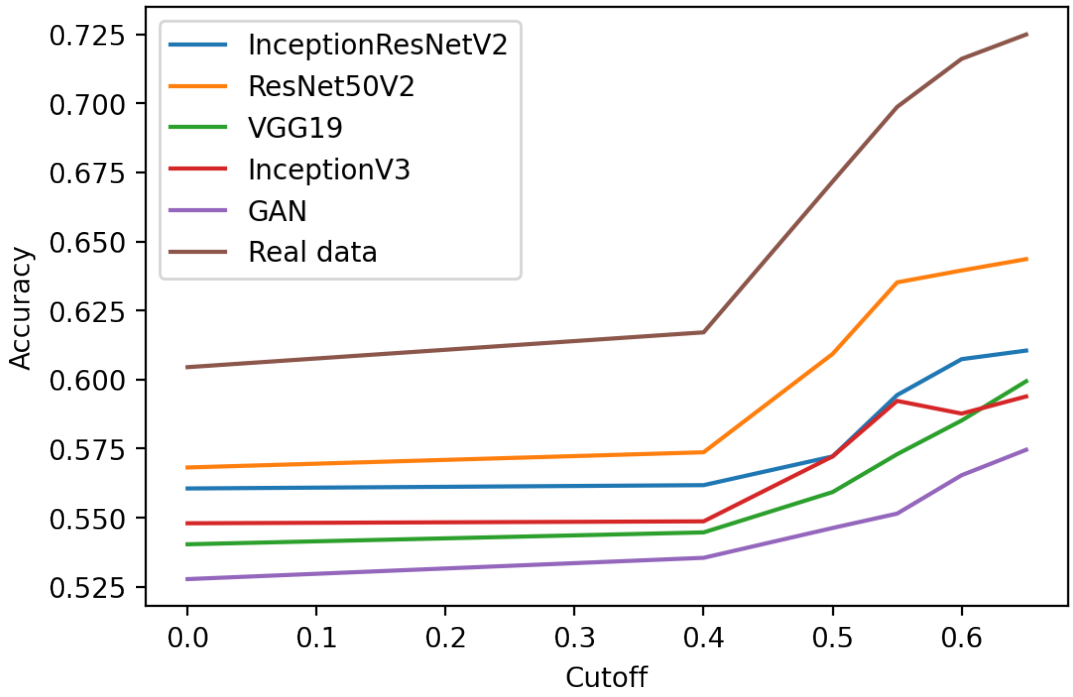


Figure 5.13: Performance comparison of classifiers for generated images for various cut-offs of different ground truths.

- Increasing the cutoff value results in higher accuracy for all ground truth measures, which supports our hypothesis of having annotators with similar perspectives in our study and ArtEmis. It also reflects that models are able to pick visual features from an painting which effect the emotion conveyed by artwork.
- For all ground truth measures, the classifiers consistently exhibit notably higher accuracies than the GAN models. This finding suggests that classifiers align more closely with human perceptual judgment when classifying emotions than GANs.
- In Chapter 4, a close competition between Inception-ResNet-V2 and ResNet50-V2 was observed. However, in human evaluations, ResNet50-V2 outperforms Inception-ResNet-V2.
- Given that StyleGAN2-ADA outperforms other GAN models regarding FID score and the Turing test, our findings further indicate that GANs are more adept at generating images with specific emotions. Although the difference in accuracy between the GAN models and the classifiers is not as significant as in other GAN models, StyleGAN2-ADA emerges as the best performer across nearly all evaluation metrics.

CHAPTER 6

Conclusions and Discussion

For the artwork generation task, StyleGAN2-ADA emerges as the top-performing model when considering a combination of overall metrics such as FID, KID, precision, and recall. It achieved the best FID score of 7.84 indicating the high similarity between the generated images and real images. For the classification task, ResNet50-V2 presented the highest accuracy of 72% for the binary emotion paradigm. This model displayed an effective balance of all metrics. ResNet50-V2 emerged as the most effective model on quantitative and qualitative metrics.

The human assessment shows the generators' ability to successfully produce affective artwork and the classifiers' ability to categorize them. Further, this work demonstrates the machine's capacity to learn and identify visual features central to eliciting specific emotions. We obtained a False Positive Rate (FPR) of 49.12% (GAN-generated images assessed as real artworks by human evaluators) and a True Positive Rate (TPR) of 77.90% (real artworks were correctly judged as human creations by the evaluators). The accuracy of the Turing test experiment was 64.03%. For all ground truth measures of generated images (generator models or classifiers), the classifiers consistently exhibit higher accuracy compared to the GAN models. This finding suggests that classifiers align more closely with human perceptual judgment when classifying emotions than GANs. In the 'Guess the Emotion' test, ResNet50-V2 outperforms all other classifiers in predicting emotion.

GANs require a large dataset for effective training. Ideally, millions of data samples are desirable. In our experiments, however, the Wikiart and ArtEmis datasets consist of thousands of artworks. Increasing the dataset size would result in a richer, higher quality dataset and facilitate better training of generative models and enabling them to capture more visual features. Consequently, this would enhance the ability of the generative models to produce high-quality artworks.

Although this research focuses on GANs, diffusion models are in vogue and demonstrably shown to be more effective than GANs. However, Diffusion models require

significantly higher computational power than what was utilized in this study.

We trained generative models on a GPU server with a single RTX 3090, with 24 GB VRAM. While we trained the GANs using approximately 10k 'kimgs' (thousands of images), it is possible to train them with a much higher number of 'kimgs' and larger batch sizes. However, this would require a high-end GPU, such as a multi-GPU server with A100 GPUs. Such an upgrade could improve the quality of generated artwork and reduce the training time. Similarly, for the classification problem, utilizing a high VRAM GPU would enable an increase in the batch size, potentially leading to improved results.

Instead of assigning a single emotion to each artwork, an alternative approach is to utilize multiple emotions for tagging a single artwork. Instead of categorizing artworks into discrete emotion classes, this approach considers a more nuanced representation by incorporating multiple emotions to capture the complex emotional aspects of the artwork.

Both generative and classification models utilize stochastic techniques, meaning they can produce varying results even with the same configuration and parameters. This stochastic nature is particularly evident in GANs, where we have observed significant deviations in FID scores, sometimes up to 7-8, even when using the same dataset and hyperparameters. These variations highlight these models' inherent randomness and sensitivity to different factors, including the initial conditions, optimization process, and training dynamics.

Assigning emotions to artworks is an inherently ambiguous task that can vary depending on an individual's gender, location, and other factors. Through our human assessment (Turing test), we observed variations in our user evaluations, with some aligning with the average annotator of ArtEmis and others deviating from it. We have engaged scholars from IIIT-Delhi as evaluators, which limits the diversity of perspectives. To address this limitation, involving a broader range of individuals, including artists, designers, historians, archaeologists, psychologists, psychiatrists, and people from different age groups, would be beneficial. By expanding the pool of annotators, we can gather a more diverse set of annotations.

In this study, we conducted Turing testing for binary emotion classes only. This can also

be extended to the multi-emotion experiment. However, for this extension, it is essential to engage professional annotators due to the nuanced similarities between emotions such as *Contentment*, *Amusement*, and *Excitement*.

.

REFERENCES

[1] S. Shahriar, "Gan computers generate arts? a survey on visual arts, music, and literary text generation using generative adversarial network," *Displays*, p. 102237, 2022.

[2] M. Edelman, *From art to politics: How artistic creations shape political conceptions*. University of Chicago Press, 1996.

[3] P. Noy and D. Noy-Sharav, "Art and emotions," *International journal of applied psychoanalytic studies*, vol. 10, no. 2, pp. 100–107, 2013.

[4] G. Strezoski, A. Shome, R. Bianchi, S. Rao, and M. Worring, "Ace: Art, color and emotion," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1053–1055.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[6] X. Wu, "Creative painting with latent diffusion models," in *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 59–80. [Online]. Available: https://aclanthology.org/2022.cai-1.8

[7] H. Zhang and G. Feng, "Enhanced example diffusion model via style perturbation," *Symmetry*, vol. 15, no. 5, p. 1074, 2023.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[9] Z.-S. Liu, W.-C. Siu, and L.-W. Wang, "Variational autoencoder for reference based image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 516–525.

[10] Z.-S. Liu, V. Kalogeiton, and M.-P. Cani, "Multiple style transfer via variational autoencoder," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2413–2417.

[11] M. Edelman, *From art to politics: How artistic creations shape political conceptions*. University of Chicago Press, 1995.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[16] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 104–12 114, 2020.

[17] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[19] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected gans converge faster," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 480–17 492, 2021.

[20] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards faster and stabilized gan training for high-fidelity few-shot image synthesis," in *International Conference on Learning Representations*, 2021.

[21] K. Dobler, F. Hübscher, J. Westphal, A. Sierra-Múnera, G. de Melo, and R. Krestel, "Art creation with multi-conditional stylegans," *arXiv preprint arXiv:2202.11777*, 2022.

[22] D. Alvarez-Melis and J. Amores, "The emotional gan: Priming adversarial generation of art with emotion," in *2017 NeurIPS Machine Learning for Creativity and Design Workshop*, 2017.

[23] "Wikiart dataset," https://archive.org/details/wikiart-dataset.

[24] "Moma. the museum of modern art (moma) collection data, 2017."

[25] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.

[26] W. Zhao, D. Zhou, X. Qiu, and W. Jiang, "Compare the performance of the models in art classification," *Plos one*, vol. 16, no. 3, p. e0248414, 2021.

[27] F. S. Khan, S. Beigpour, J. Van de Weijer, and M. Felsberg, "Painting-91: a large scale database for computational painting categorization," *Machine Vision and Applications*, vol. 25, pp. 1385–1397, 2014.

[28] S. Bianco, D. Mazzini, P. Napoletano, and R. Schettini, "Multitask painting categorization by deep multibranch neural network," *Expert Systems with Applications*, vol. 135, pp. 90–101, 2019.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[31] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[32] S. Aslan, G. Castellano, V. Digeno, G. Migailo, R. Scaringi, and G. Vessio, "Recognizing the emotions evoked by artworks through visual features and knowledge graph-embeddings," in *Image Analysis and Processing. ICIAP 2022 Workshops: ICIAP International Workshops, Lecce, Italy, May 23–27, 2022, Revised Selected Papers, Part I*. Springer, 2022, pp. 129–140.

[33] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "ArtEmis: Affective language for visual art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 569–11 579.

[34] Y. Mohamed, M. Abdelfattah, S. Alhuwaider, F. Li, X. Zhang, K. W. Church, and M. Elhoseiny, "ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture," *arXiv preprint arXiv:2211.10780*, 2022.

[35] Y. Mohamed, F. F. Khan, K. Haydarov, and M. Elhoseiny, "It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 263–21 272.

[36] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[37] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[38] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.

[39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[43] 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.

[44] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[45] A. M. Turing, *Computing machinery and intelligence*. Springer, 2009.