Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Classification of Real images from DeepFakes

Ashutosh Sharma, Mrs. Anubhooti Papola
Veer Madho Singh Bhandari Uttarakhand Technical University, Dehradun

Abstract: The progress of Artificial Intelligence has brought along certain disadvantages primarily generation of Deepfakes that are being circulated through various channels of communication. These images are impossible to differentiate from real images due to their exceptional quality. This project is aimed at classifying real images from AI generated images (generated using diffusion models). This paper will draw a comparative study between various standard classifiers such as Gradient Boosting, K Nearest Neighbor, Random forest and Neural Network to find out which algorithm is best suited for this type of classification task.

1. Introduction

The study evaluates the performance of the standard classifiers in classifying the real images from AI-generated images. To carry out the study, first we need to understand in depth about the problems that can be created using AI-generated images.

1.1 History of Information Manipulation

People have always been misguided by spreading misinformation. There is a long list of scenarios that represents the circulation of false news. The article "Great Moon Hoax" published in Sun newspaper of New York in 1835 claimed that there existed an alien civilization on the moon, making it the most read newspaper in New York [1]. An attempt to gain audience by using fake news was made by American publishers Joseph Pulitzer and William Hearst in the 1890s. The misinformation spread through newspapers took week to circulate and had the possibility of being withdrawn before it could reach a larger set of people. However, in this time of global connectivity spreading misinformation has become child's play. Anyone with access to internet can generate images enacting reality using text based prompts

.

1.2 Brief Understanding of GANs

The introduction of GANs (Generative Adversarial Networks) in 2014 marks the first step in practical image generation. GANs consist of two main blocks:

- **Generator**

This block takes noise as an input and works to generate images that are close in identity to those in the original dataset. It tries to learn P (X|Y) which is the joint probability of input data X and output data Y.

- **Discriminator**

This block takes two inputs, one from the original dataset and other from the generator and then classifies the image as legitimate or deepfake.

As the learning process continues, the generator improves its ability of generating data while discriminator refines its bifurcation ability.

The following figure showcases the results of Yu and Porikli's work.

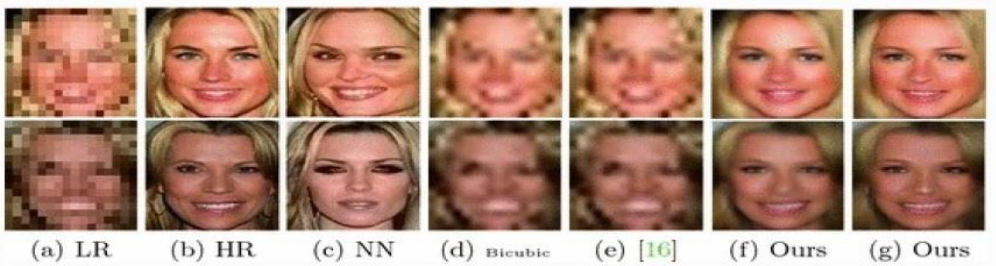


Fig 1: Faces generated using GAN by Yu &Porikli

1.3 Limitation of Image Generated by GAN

GANs employ deconvolutional layers that may sometimes generate images with artifacts, distortions, or undesired patterns. It is observed that images generated using GANs are prone to checkerboard artifacts i.e. the generated images contain visual patterns resembling a checkerboard [2]. These patterns in generated images are undesirable as they degrade the visual quality of generated images, make them less realistic and easy to classify from the real images. .

1.4 Introduction to Diffusion Model

With GAN architecture being prone to deconvolution and checkerboard artifacts, there was need for a generative AI model that could generate high quality realistic images. This led to the introduction of diffusion models. These models employ a forward and a reverse diffusion process in which they gradually add Gaussian noise to the original data and then remove the noise respectively. OpenAI has created one of the most popular diffusion models, named Dall-E 2. This AI system uses a natural language description to generate realistic images and art.

Figure 2 showcases high quality, diverse and realistic images created using Dall-E 2

Figure 2: Results from Dalle-2

1.5 Motivation

With the introduction and easy availability of power of image generation tools such as the diffusion models, it is possible for any individual having computing power to generate realistic images by providing vivid description. The main demerit of this power is that it can be used to generate any picture that may show an individual doing things that he/she might have never done. Images and videos generated through these AI generative tools can and will have negative consequences. Media generated can be used to humiliate, attack individuals as well as to instill fear and anger among masses. There have been events in past where such generated media has been used to trigger a large scale protest. There is a need for a classification tool that can bifurcate the real images from the generated ones.Specifically, the goal of this project is to find out which classifying technique is best suited for the task.

1.6 Objective

The aim of this research work is to calculate the accuracy of various classifiers in classifying the real images from the AI generated images (generated using diffusion models) and find out the classifier with the highest accuracy in above classification process.

2. Literature Review

This chapter discusses about research works that have already been completed in relation to identification of AI generated images. Since image generation through diffusion models is very recent, most of the research that has been carried out pertains to Generative Adversarial Networks. There are only a handful of papers discussing diffusion models.

In 2022, the Advances in Computer Vision and Patter Recognition book series consisted of a handbook that was focussed on “Digital Face Manipulation and Detection” [3]. The handbook aimed to provide inclusive overview of generated image detection. The various techniques that were discussed included color features analysis, identifying asymmetries in caseand using data-driven features.

Researchers observed thatGAN fingerprint detection method is extremely effective in identification of generated images. Zhang et al. showcased that the GAN pipeline left a distinctive checkerboard artifact in the upsampling stage which acts like a unique fingerprint in each image. This unique fingerprint can not only be used to identify fake images but also to determine the model responsible for this fake image [4]. The GAN image classifier achieved a classification rate of 0.95 when with spectral inputs.

Figure 3 showcases an example of identifying the distinctive artifact left behind in generated image.

Fig 3: Zhang et al.’s artifact detection

Roy and co-authors tested variety of CNN architectures and models for classification an detection of deepfakes. Using ResNet in combination with 93,647,040 trainable parameters they achieved a classification rate of .9550 [3].

Meanwhile, researchers also looked at other visual artifacts such as blurriness and face warping to train classifiers in order to identify deepfakes. These research works demonstrate that it is possible to classify images generated using GANs and also achieving a high classification rate while doing so.

As already stated, the topic of GAN image detection has been thoroughly researched but because the images generated by diffusion models like Dall-E 2, does not have the same shortcomings as in GAN, the process of classification is complex. The authors of the paper that describes the creation process and working structure of Dall-E 2 state, "Gaussian blur has been employed for the first upsampling stage whereas a more diverse BSR degradation has been used for the second" [5]. Zhang et al developed this upsampling layer that involves generating noise through the addition of Gaussian noise with varying intensities, adopting JPEG compression with different quality factors, creating simulated noise patterns that resemble the noise introduced by camera sensors during image capture [6]. Figure 4 shows how a degraded image can be refined by applying upsampling architecture.

Fig 4: Image upsampling as depicted by Zhang et al. depicting

The methods utilized for classifying GAN does not work well with its counterpart because of lack of checkerboard artifacts in diffusion models.

"Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models" study undertook by Sha et al. In January 2023 also focussed on diffusion models such as Dalle-2 and Stable diffusion [7]. MSCOCO and Flickr30k datasets were used to conduct experiments by researchers. These datasets comprises of real images along with textual information describing these images. Models such as Dall-E 2 were used to generate images based on the provided textual information with an aim to train classification models for distinguishing between "real" or "fake" image. The below figure showcases the accuracy scores achieved in the aforementioned research with respect to different diffusion models.

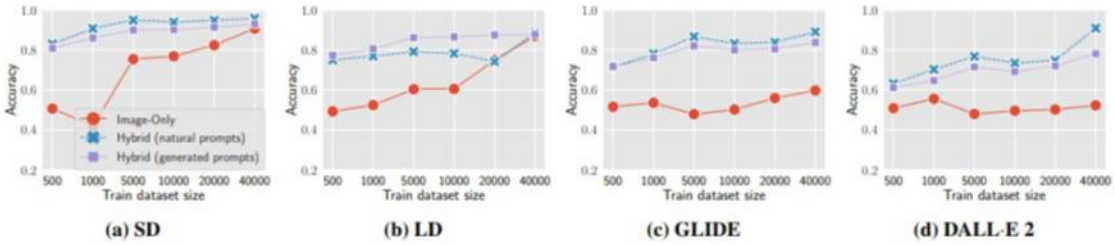


Fig 5: Classification scores achieved by SD, LD, GLIDE and Dall-E 2 [8]

The classification score for latent diffusion and glide models was approximately 0.9 and for Dall-E 2 and stable diffusion was approximately 0.95. One important point to consider, in this classification task, is that the model uses information about both prompt as well as the image, therefore, it can be inferred that the classification accuracies achieved will be limited when datasets containing solely the images and prompts will be provided to the classifier. That being said, still this research demonstrates that diffusion models exhibit informative features that can be exploited for classification.

Though the research into the classification of images generated using diffusion models is limited, still they offer evidence that such classification is achievable. In conclusion, it can be said that there is a shortcoming in detecting images generated from these models as well. This project aims to draw a comparative analysis between various standard classification algorithms by comparing their classification accuracies in bifurcating Real and AI generated images (generated using diffusion models).

3. Performance Evaluation Metrics

The ability of our classifiers is estimated by using various performance evaluation metrics which are discussed below:

3.1 Confusion Matrix

In machine learning and statistics, confusion matrix (also called error matrix) is used to estimate the performance of classification models. It displays summarized picture of classifications made by the classifier compared to the actual class labels. This helps in carrying out a thorough analysis of the classifier's performance. It provides a summarized view of the model's performance.It helps finderror in classification result and used as a basis to derive values of other important metrics namely precision, recall, F1-score, and accuracy.

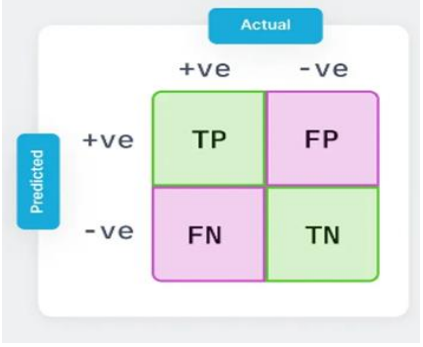


Fig 5: Graphical Representation of Confusion Matrix

3.2 Accuracy

Accuracy is used for examining the overall performance of a classification model. It is defined as the ratio of the correctly predicted instances to the total instances in the dataset. It is calculated as follows:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Accuracy is the fraction of correct predictions out of the total predictions generated by the model. Accuracy is useful when the class distribution is balanced in nature. It is not advisable to use accuracy as performance evaluation metric when dataset contains imbalanced class distribution.

3.3 Precision

Precision (denoted as P) is the ratio of true positive predictions to all positive predictions. It tells us that how accurately the model has made the positive predictions. Higher the value of precision, lower the number of false positives. Its formula is:

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

3.4 Recall

Recall (denoted as R) is also known as true positive rate (TPR) or sensitivity, is the ratio of true positive predictions to all actual positive instances. Higher the value of recall, lower the number of false negatives. It shows the ability of the classifier to correctly identify positive instances. The formula for recall is:

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

3.5 F1 Score

It is calculated using the values of precision and recall by taking the harmonic mean of both. It is beneficial when dealing with datasets that contain an imbalanced class distribution. The formula for the F1 score is:

$$\textbf{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.6 Receiver Operating Characteristic Curve

It is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate or Recall

$$\textbf{TPR} = \frac{TP}{TP + FN}$$

- False Positive Rate

$$\textbf{FPR} = \frac{FP}{FP + TN}$$

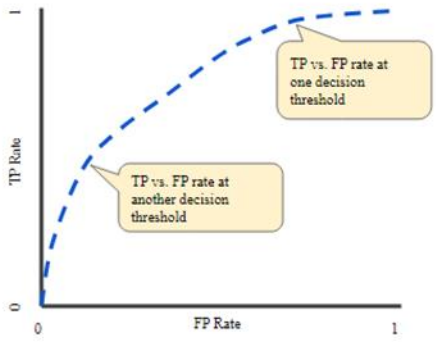


Fig 6: Graphical Representation of ROC Curve

3.7 Area under the ROC Curve

AUC measures the entire two-dimensional area underneath the entire ROC curve from (0, 0) to (1, 1).

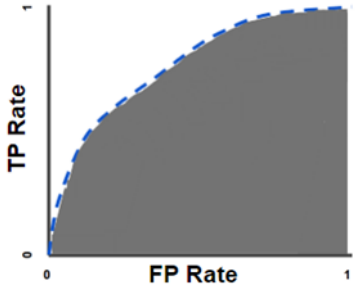


Fig 7: Graphical Representation of AUC in ROC Curve

AUC has value in range of 0 to 1. An AUC score of 0.0 indicated that all the predictions made by the model are incorrect whereas an AUC score of 1.0 indicated that the models have made all predictions correctly.

3.8 MCC

MCC is the summary of the confusion matrix (error matrix) that is considered to be the best single-value metric. A confusion matrix comprises of: True positives (TP), True negatives (TN), False positives (FP) and False negatives (FN).

$$\mathbf{MCC} = \frac{\mathbf{TN \times TP - FN \times FP}}{\sqrt{((\boldsymbol{TP}+\boldsymbol{FP})(\boldsymbol{TP}+\boldsymbol{FN})(\boldsymbol{TN}+\boldsymbol{FP})(\boldsymbol{TN}+\boldsymbol{FN}))}}$$

If the prediction returns good rates for all four of these entities, it is said to be a reliable measure producing high scores. And to suit most correlation coefficients, MCC also ranges between +1 and -1 as:

- +1 is the best agreement between the predicted and actual values.
- 0 is no agreement. Meaning, prediction is random according to the actual

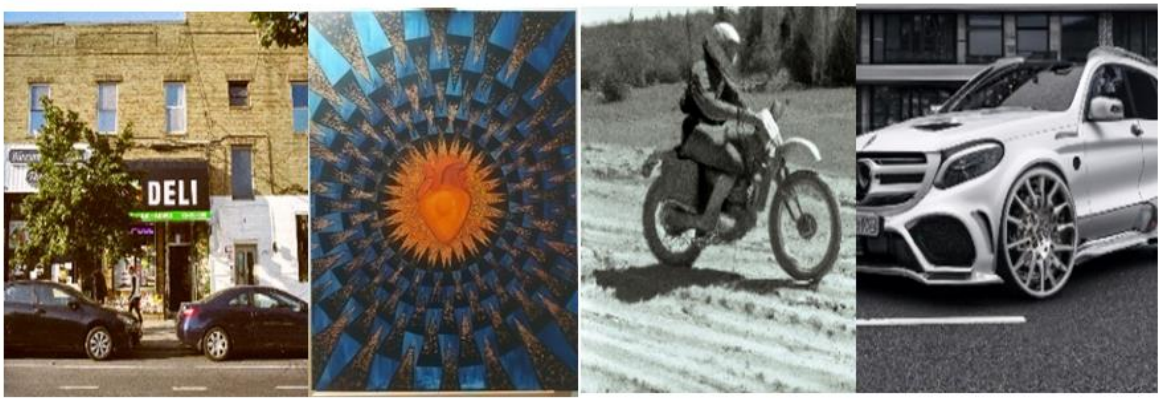
4. Implementation

Implementation of the work involves various steps and resources which are detailed as follows.

4.1 Dataset

The most important part of any research is the dataset used in order to carry out the research.

The dataset selected is diverse in nature, containing ample number of real and AI generated images. During literature survey, it was observed that previous work, with respect to diffusion models, made use of a combination of real objects and prompts associated with these objects for generating images. This work, on the other hand makes use of publically available images generated by users using different diffusion models. The dataset contains around 3500 "real" images collected through various sources such as photographs, digital art, movies, user-drawn and portraits. When compared to the MSCOCO dataset that is often used in previous researches, here "Real" and "AI-Generated" image classes not entirely contain generic images (tree, bag, and cat). The "AI-Generated" class contains images that are generated by different users using different diffusion models. In similar manner, the "Real" class also contains digital art.



(a) (b) (c) (d)

Fig 8: Images (a) and (b) are Real whereas (c) and (d) are AI-generated

4.2 Methodology Used

The project aims to draw a comparative analysis between various classification algorithms to find out the algorithm that performs the best in out scenario. The classification algorithms used in the project are:

4.2.1 Gradient Boosting

It is a machine learning algorithm that is used for classification and regression task. It is a popular boosting algorithm that is a kind of ensemble learning method in which the model is trained sequentially where each new model tries to improve the previous model.

4.2.2 Random Forest

It is one of the most popular supervised machine learning algorithms. In similar manner to Gradient Boosting, Random Forest can be used for both Classification and Regression tasks. Its working utilizes the concept of ensemble learning, where multiple classifiers are combined with an aim to solve a complex problem as well as to improve the performance of the model.This classification algorithm makes use of a number of decision trees on various subsets of the given dataset and takes the average to improve the accuracy of prediction of that dataset**.** Random Forest predicts the final output on basis of majority votes of predictions taking the prediction from each tree into consideration.

4.2.3 K- Nearest Neighbor

K-Nearest Neighbour is one of simplest supervised machine learning algorithm. It divides the data into various categories and when new data is presented to the algorithm, it classifies the data into one of the available categories depending upon its similarity with that category. Like Gradient Boosting and Random Forest, KNN can also be used for both classification as well as Regression tasks. However, it is mostly utilized in classification of data. KNN algorithm during the training phase just stores the dataset and when it gets new data, and then it classifies that data into a category that is much similar to the new data. As a result it is also termed as Lazy Learner algorithm.

4.2.4 Neural Network

A neural network is a machine learning model that imitates the structure and function of a human brain. In neural networks, nodes (also referred to as neurons) are interconnected in a complicated manner. These nodes

combine to solve complicated problems. It has diverse uses such as image recognition, predictive modeling and natural language processing (NLP).

4.2.4.1 Stochastic Gradient Descent

It is utilized for optimizing machine learning models. It overcomes the inability of traditional Gradient Descent algorithms in processing large datasets in machine learning projects. SGD overcomes this problem by selecting a random training batch from the dataset to calculate the gradient and update the model's parameters instead of using the entire dataset for each iteration.

4.2.4.2 ADAM

Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data.Adam was introduced by Diederik Kingma from OpenAI and Jimmy Ba from the University of Toronto in their 2015 ICLR paper titled "Adam: A Method for Stochastic Optimization" [8].

4.3 Algorithm

The dataset discussed above contains both real as well as AI- generated images. The dataset is divided into two parts for training and testing phases. The training set comprises of 80 % images in the dataset and the testing set comprises of remaining 20% images. The images are then fed to different standard classifiers such as KNN to evaluate their performance in the training as well as testing phase. Later, a comparative study is drawn to determine which classifier has outperformed all others in terms of classification accuracy.

4.3 Approach

The working of the project is explained in a simplified manner using the following figure:

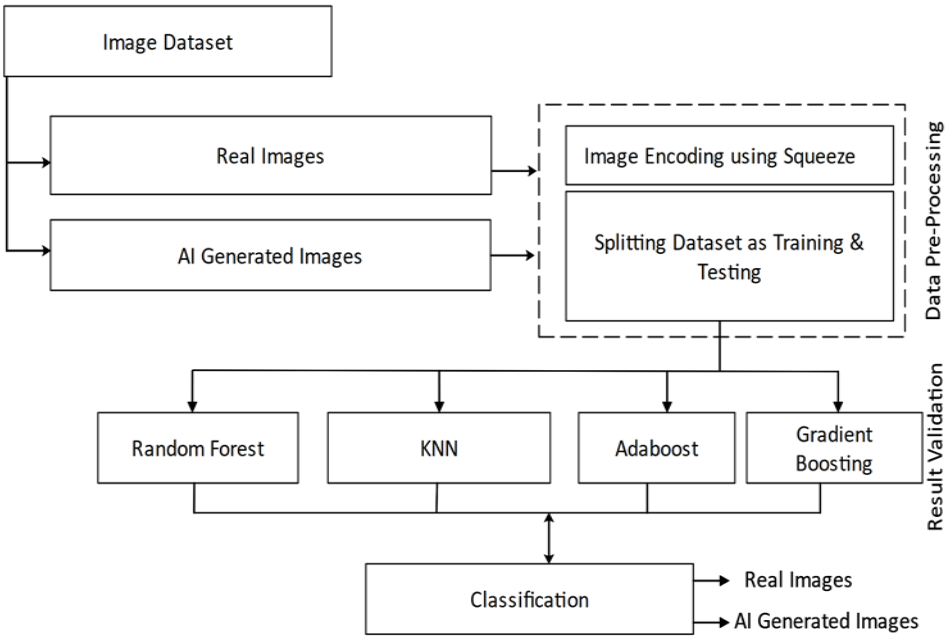


Fig 9: Research work flow diagram

5. Result

The classification algorithms were evaluated separately for their performance during training and testing process. To decide which system has the best performance quantitatively, we have made use of accuracy score. Since the dataset contains a balanced distribution of classes between the labels, accuracy is chosen as the performance metric to compare the classification models.We made use of different classification algorithms

such as Gradient Boosting, K- Nearest Neighbor, Random Forest, Neural Network (using SGD as optimization algorithm) and Neural Network (using ADAM as optimization algorithm), it is observed that out of all the algorithms used to classify real images from AI-generated images the Neural Network that uses SGD as optimization algorithm achieves the highest classification score during training as well as testing phase. The training accuracy score for SGD based neural network was 0.848 and the testing accuracy score was 0.851.

S.No	Classification Approach	Classification Score
1	Gradient Boosting	0.817
2	KNN	0.800
3	RF	0.765
4	Neural Network (ADAM)	0.840
5	Neural Network (SGD)	0.848

Table i: Classification scores achieved by various approaches during training phase

S.No	Classification Approach	Classification Score
1	Gradient Boosting	0.815
2	KNN	0.817
3	RF	0.766
4	Neural Network (ADAM)	0.822
5	Neural Network (SGD)	0.851

Table ii: Classification scores achieved by various approaches during testing phase

PREDICTED

ACTUAL

	Dall-E Samples	Real	∑
Dall-E Samples	1331	481	**1812**
Real	334	2303	**2637**
∑	**1665**	**2784**	**4449**

(A)

PREDICTED

ACTUAL

	Dall-E Samples	Real	∑
Dall-E Samples	1319	493	**1812**
Real	398	2239	**2637**
∑	**1717**	**2732**	**4449**

(B)

PREDICTED

ACTUAL

	Dall-E Samples	Real	∑
Dall-E Samples	1281	531	**1812**
Real	515	2122	**2637**
∑	**1665**	**2784**	**4449**

(C)

PREDICTED

ACTUAL

	Dall-E Samples	Real	∑
Dall-E Samples	1331	481	**1812**
Real	334	2303	**2637**
∑	**1665**	**2784**	**4449**

(D)

Fig 10: Confusion Matrix of (A) Gradient Boosting, (B) KNN, (C) Random Forest and (D) Neural Network (Training Phase)

PREDICTED

ACTUAL	Dall-E Samples	Real	∑
Dall-E Samples	550	213	**763**
Real	139	1004	**1143**
∑	**689**	**1217**	**1906**

(A)

PREDICTED

ACTUAL	Dall-E Samples	Real	∑
Dall-E Samples	566	197	**763**
Real	151	992	**1143**
∑	**689**	**1217**	**1906**

(B)

PREDICTED

ACTUAL	Dall-E Samples	Real	∑
Dall-E Samples	524	239	**763**
Real	207	936	**1143**
∑	**689**	**1217**	**1906**

(C)

PREDICTED

ACTUAL	Dall-E Samples	Real	∑
Dall-E Samples	611	152	**763**
Real	118	1025	**1143**
∑	**689**	**1217**	**1906**

(D)

Fig 11: Confusion Matrix of (A) Gradient Boosting, (B) KNN, (C) Random Forest and (D) Neural Network (Testing Phase)

6. Conclusion and Future Prospect

After carrying out the extensive study, the following conclusion is drawn. In addition to the conclusion, some future advancement regarding the research are also presented in detail.

6.1 Conclusion

This idea behind the project is to identify the images generated using diffusion models such as Dall-E 2. As the field of generative AI becomes more and more advanced, it has become very difficult to tell a deepfake apart from a real image. It is very important for news networks, social media companies as well as users to carefully evaluate every bit of media (image or video) to prevent any harm that it could cause. The project tries to draw a comparative analysis between different classification techniques such as Gradient Boosting, Random Forest, KNN, and Neural Network to find out the technique that is best suited for this particular classification. Based on the accuracy score, SGD turns out to be the most accurate in performing this classification task. This project has made way for further research in Generative AI especially related to the identification of deepfakes generated using diffusion models.

The model was provided two different images (one real and one AI-generated) of Uttarakhand Technical University. Upon classification, the model successfully classified the real and AI generated image. The results were as follows:

(a) AI generated image successfully classified as AI generated

(b) Real image successfully classified as Real

Fig 12: Classification results

6.2 Future Work

This project has tremendous future prospects. It can be improved in many different ways such as using a larger dataset to train the model, using better optimization algorithms than the ones taken into consideration or by using more complex neural network architecture such as an ensemble approach to better classify the data. Any research work carried out on any of the above prospects can yield much better classification accuracy.

7. Conflict of Interest Statement

There is no conflict of interest for publication of this manuscript.

References

[1] J. Soll, “The Long and Brutal History of Fake News,” POLITICO Magazine, Dec. 18, 2016. http://politi.co/2FaV5W9 .

[2]Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10), e3.

[3] Christian Rathgeb et al. Handbook of digital face manipulation and detection: From deepfakes to morphing attacks. Springer Nature, 2022.

[4] Zhang, X., Karaman, S., & Chang, S. F. (2019, December). Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.

[5] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

[6] Zhang, K., Liang, J., Van Gool, L., & Timofte, R. (2021). Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4791-4800).

[7] Sha, Z., Li, Z., Yu, N., & Zhang, Y. (2023, November). De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (pp. 3418-3432).

[8]Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.