Deepfake Detection: A Comparative Analysis of Deep Learning Algorithms

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Prashanthi Thota- (AP21110010069)

KPVM Karthik- (AP21110010263)

Sashank Sakkurthi- (AP21110010272)

Lakshmi Nikhitha Dodda- (AP21110011270)

Under the Guidance of

Dr. Murali Krishna Enduri

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[Nov, 2024]

Certificate

Date: 14-Nov-24

This is to certify that the work present in this Project entitled "**Deepfake Detection: A Comparative Analysis of Deep Learning Algorithms**" has been carried out by **Prashanthi Thota, KPVM Karthik, Sashank Sakkurthi, Lakshmi Nikhitha Dodda** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

Dr. Murali Krishna Enduri

Assistant Professor, Head of the Department (CSE),

Computer Science and Engineering.

Co-Supervisor

Dr. Ravi Kant Kumar

Assistant Professor,

Computer Science and Engineering.

Acknowledgements

We extend our sincere gratitude to everyone who contributed to the successful completion of our Deepfake Detection: A Comparative Analysis of Deep Learning Algorithms project. Throughout this journey, we have been fortunate to receive continuous support, valuable advice, and unwavering encouragement.

A special note of appreciation goes to Dr. Murali Krishna Enduri whose extensive knowledge and support played a crucial role in shaping the course of this Interdisciplinary project.

Looking back, being part of this challenging yet fulfilling project has been an honor. We are eager to apply the skills and knowledge gained to further advance our careers in Computer Science.

Prashanthi Thota (AP21110010069)

KPVM Karthik (AP21110010263)

Sashank Sakkurthi (AP21110010272)

Lakshmi Nikhitha Dodda (AP21110011270)

Abstract

In the contemporary world of high technology, deepfake is now at once a technological marvel and a concern. Based on deep learning, deepfakes make fakely sophisticated media, which is often hardly distinguishable from the unrealised one. However, novel deepfake technology raises critical ethical, security, and privacy issues and risks in areas including misinformation, cybercrime, and identity theft. Ensuring the authenticity of content has therefore turned out to be an important mission in combating fakes. Currently, deep learning involves the identification of deep fakes since it is capable of understanding dataset characteristics that are complex to human comprehension. There are convolutional neural networks (CNNs), VGGNet, ResNet, and AlexNet; they are used in hierarchy feature extraction to detect fake media. Such representations enable the detection of fine-grained features in images that are different from genuine and create artifacts peculiar to deepfake content. On the other hand, for realistic media synthesis, generative adversarial networks (GANs) were expanded for detection tasks by trying to detect synthetic content based on the adversarial structure of the generator-discriminator. This research project presents a comparison of CNN and GAN structures to understand which approach works better and why when it comes to deepfake detection. CNNs such as VGGNet and ResNet provide high accuracy and relatively equal characteristics in the detection of various types of manipulations. However, comparatively lighter models, including AlexNet and GAN-based discriminators, which involve low computational cost, could be employed as an ensemble. Stressing the need to develop and apply methods for deepfake detection, this study shows how modern deep learning methods can help to ensure security, provenance, and quality of digital media in the era of synthetic environments.

Index Terms: Deepfake Detection, Deep Learning, CNN-VGGNet, CNN-ResNet, CNN-AlexNet, and Generative Adversarial Networks (GANs).

Table of Contents

Abbreviations

CNN - Convolutional Neural Network

GAN - Generative Adversarial Network

VGGNet - Visual Geometry Group Network

ResNet - Residual Network

List of Tables

List of Figures

List of Equations

1. Introduction

Deepfake technology based on hyper-real manipulative media became the new brilliant creation of machine learning and, at the same time, a new social threat. Creating from entertainment and artistic points of view, deepfake features have expanded to spheres with vital moral consequences, such as fake news, stealing one's identity, or cyberspying. There is a high possibility of creating deepfake content with pictures, voices, and videos interweaved with realistic versions intertwined with original fakes to circumvent human as well as artificial intelligence detections. Realization of such technology has grown at a very fast pace, thereby placing much demand on efficient methods of detecting deep fakes. At the heart of this challenge is one of the most powerful tools in the machine learning tool chest, deep learning, which has shown capability in the establishment of detection of fake media and origins from synthetic ones. This paper adopts strict comparative analysis of several of the most standard deep learning models, including CNN-VGGNet, CNN-ResNet, CNN-AlexNet, and GAN, with each having its own strengths and weaknesses in deep fake technology. The main challenge of deepfake as a field, especially in terms of developing reliable detection methods, is the exponential progress of deep learning approaches in the creation of the media. While in forgery of traditional media, irregularities might be detected by the naked eye or simple mathematical algorithms, deepfakes are best distinguished by methods that can detect subtleties in patterns, texture, facial expressions, and sound quality. As the instances of deepfake applications rise, these detection models need to work efficiently on various datasets and use cases and still be immune to attacks that aim for evading detection.

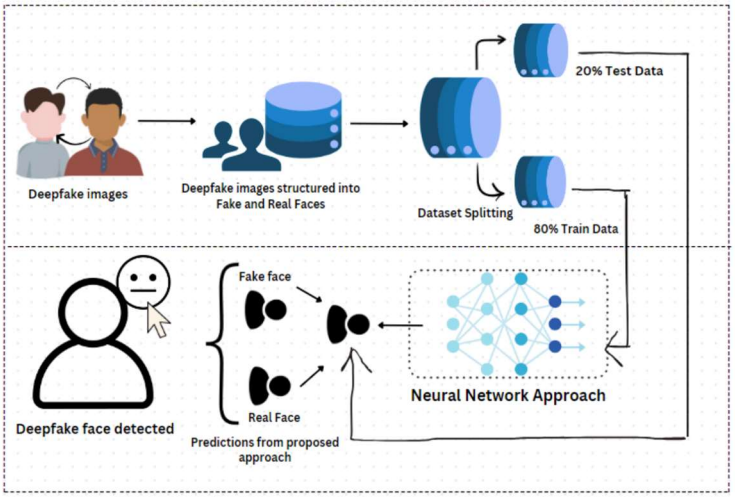

Fig. 1: Deepfake detection workflow with dataset splitting and neural network processing.

1.1 The Role of Deep Learning in Deepfake Detection: Research on image, audio and video analysis is mainly driven by deep learning due to its ability to discover complex features of data without requiring prior knowledge. In detail, Convolutional neural networks (CNNs) have shown high efficiency in carrying out visual recognition tasks and, therefore, are the most preferred networks for identifying fake images and videos. CNNs rely on down-and-up sampling, where features are hierarchically built, enabling the technique to detect subtle features specific to deepfake content. However, to capture deepfakes, one needs a model that is not only good but robust because new nefarious synthesis methods are always being developed. Due to their flexibility for solving intricate visual problems, CNNs have become very useful in detecting deepfake. This is because it is a never-ceasing game of designing models to be capable of countering the ever-emerging challenges. As for generative adversarial networks (GANs), for example, the technique opens an all-new dimension of detection due to the dual structures of the generator and discriminator. Although initially applied to realistic media synthesis, GANs have the possibility of detection tools when configuring the discriminator for recognising discrepancies between synthesis and actual media. This adversarial framework enables a smaller number of false positives but is computationally expensive. It is the focus of this work to compare the effectiveness of CNN and GAN-based models to identify when each is at its best depending on evaluation measures and database.

1.2 Deep Learning Models and Connection with Deepfake Detection: This research focuses on the comparative performance of four prominent deep learning architectures: for instance, VGGnet and Resnet, Alexnet, GAN, etc. Both models represent different architectural concepts for addressing the issue of image recognition; while CNN is tailored on a hierarchy model, GAN offers a solution based on adversarial training.

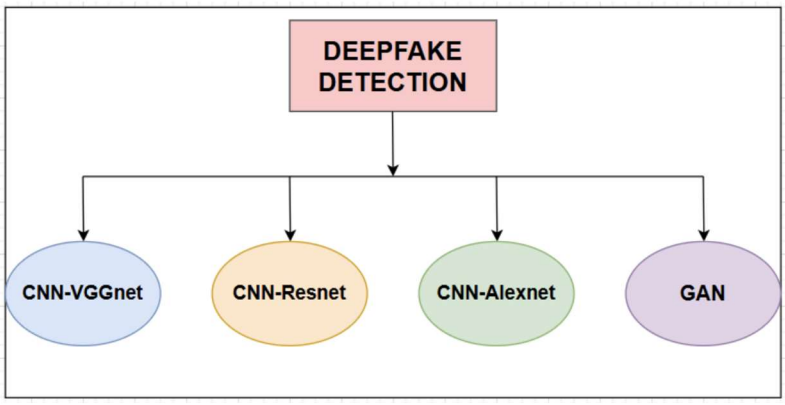

Fig. 2: Algorithms used in this study to detect deep faked images.

1.2.1 CNN-VGGNet: The Visual Geometry Group (VGG) developed VGGNet as a type of deep CNN, which was famous for its simplicity and depth. Composed of multiple 3x3 convolutional layers connected and stacked deeper, VGGNet shows high accuracy in detecting patterns in images, essential for detecting small artifacts inherent in deepfakes. The ability of VGGNet to extract deep features for classification has made its work a strong baseline for most visual recognition tasks, including deepfake detection.

1.2.2 CNN-ResNet: ResNet presents an idea of short connections called residual connections, which help to train deep neural networks to learn identity mappings that help to overcome one of the critical problems in convolutional networks, the vanishing gradient problem. Thus, the architecture under consideration is characterized by the possibility of correspondingly high accuracy while minimizing the risks of overfitting. ResNet's residual blocks enhance the detection of layered features and thereby are promising for detecting the more intricate manipulations involved in deepfake media. Due to the high degree of efficiency and accuracy, ResNet has become the model of choice in many of the critical visual recognition applications.

1.2.3 CNN-AlexNet: Originally published in 2012, AlexNet is a deep CNN architecture with strongly associated alternating convolutional and pooling layers meant to capture spatial hierarchies in images that can help the networks succeed in computer vision tasks. However, as has been mentioned, it is slightly simpler than the VGGNet and ResNet and still due to the fact that it does not take long to train and recognises images quite effectively, so it is still quite a good option for the deepfake detection. The relatively shallow design of AlexNet may be beneficial for datasets for which speed is a more relevant criterion than depth because it presents a reasonable trade-off between depth and performance.

1.2.4 GAN: Discriminator from generative adversarial networks is especially chosen for deepfake generation as well as detection because of the adversarial training of the models. The discriminator, which to date has been used to replicate realistic fake data sets within a GAN, can be employed to also identify fake data by identifying adversarial signals. The iterative adversarial process enhances the discriminator to minimize small discrepancies undetectable by other CNN-based models, although much more time-consuming.

2. Related Work

The rapid progress of deepfake technology has made tremendous efforts possible on effective ways of detecting such manipulated artifacts, especially deep learning-based approaches. CNNs and GANs are two streams that have been broadly researched as tools for detecting manipulated media. A review on face manipulation and deepfake detection discussed the suitability of CNNs and GANs for the identification of low-level artifacts, including lighting and texture inconsistencies and facial geometry variations. This survey, too, emphasized the need for developing advanced techniques to counter the changing nature of deepfake content.

Critical algorithms and their applications in computer vision have been discussed in detail to analyze and detect deepfakes. Algorithms work on hierarchical feature extraction and advanced visual recognition methodologies that prevent such small anomalies from surfacing in synthetic media. Digital image forensics have also been utilized to thwart deepfakes, and scientists are focused on the techniques of detecting edited content by observing spatial and temporal inconsistency, as well as those image-level artifacts. These techniques have played a significant role in unearthing the flaws of the currently developed detection technologies and paving the way for robust solutions.

More recent research even provides numerical estimates to describe what is happening in the sphere of deepfake detection today by emphasizing the practical constraints in the methodology and the challenges faced in reliable detection in real life. What is needed is scalable, efficient, and accurate models in detection against increasingly sophisticated techniques used in deepfakes.

By basing it on these foundational works, the study will strive to create a comparative analysis of the leading deep learning models applied, such as VGGNet, ResNet, AlexNet, and GANs, in order to review the strengths and limitations of such architectures. The insights developed contribute to the development of robust and scalable deepfake detection systems and can help solve the growing threat associated with synthetic media.

3. Methodology

The study employs four deep learning architectures, namely VGGNet, ResNet, AlexNet, and GAN, and tests for the performance of each in detecting deepfakes against benchmark performance metrics. Both models are trained and tested on a set of real and fake (deepfake) images, and the authors focus on how these architectures perform when dealing with specific issues of deepfake detection. Further, the given section describes the method used to implement each model for further details regarding the procedure.

1. CNN-VGGNet:

Preparation and processing of the dataset: On the dataset side, load the dataset and include both real and fake images. We resize these images such that they fit VGGNet's input layer at 224x224 in size. Dividing pixel values by 255 normalizes pixel values to the range [0, 1] and would stabilize training and help the model better converge. Apply data augmentation in order to improve generalizability. Overfitting is prevented by using techniques such as rotation, width shift, zoom, etc.

Training Methodology: The base model initializes with VGG16, already pre-trained with ImageNet. Freeze all the layers except the last 4 to prevent the base model from learning general feature extraction patterns, and the last 4 can be trained to learn deepfake specific patterns. Add fully connected layers: To reduce the spatial dimensions, we use the Global Average Pooling Layer. Non-linearity using dense layers with ReLU activation. Regularization by dropout and batch normalization layers is used to regularize the model and improve generalisability. We compile the model with a categorical cross-entropy loss function and Adam optimiser with a learning rate of 0.0001.

2. CNN-ResNet:

Data Acquisition and Data Cleaning: Clean the dataset to ensure they meet ResNet specifications by resizing the images to 224*224 pixels. Standardize images to further the stability of the model and also bring convergence at an earlier time. Increase data by, similar to VGGNet, rotating, zooming, flip and other similar types of operations.

Training Methodology: ResNet50 is selected, and the pre-trained model is ResNet50, which is also initialized with ImageNet weights. Ties all but the top ten layers so as to maintain overall appearance and fine-tune the constructed model on deepfake detection. Add additional fully connected layers: This is followed by the Global Average Pooling Layer, then followed by the dense layers incorporating batch

normalization and dropout. Select categorical cross entropy for regression; optimize as Adam at a rate of 0.0001.

3. CNN-AlexNet:

Data collection and data preparation: All images should be resized to fit 227 such that they meet the input size for the AlexNet model. Within the pixel values, standardize them, and then put through data augmentation as was done with other CNN models above.

Training Methodology: In order to implement AlexNet, we start with the characteristic of having in the network successive layers of convolution and pooling. Design additional fully connected layers: Add two convolutional layers with dropout layers in between in order to mitigate too much overlearning. Note that for non-linearity, ReLU activation functions should be used and the final softmax for classification only. Train the model with categorical cross entropy and using the Adam optimiser at a learning rate of 0.0001.

4. GAN:

Dataset preparation and processing: Resize the images to fit a 64 x 64 input frame that is the architecture of the GAN. Then, resize images to 64x64 and normalize the pixel ranges to [-1, 1], which is the output range of the generator's tanh activation function.

Training Methodology: Construct a generator model with layers in order to generate synthetic realistic images from random noise. Train the discriminator model to recognise real and fake images to decide which it is dealing with. Train the GAN by: This is achieved by feeding the discriminator with real images and generated images to minimize the parameter update of the discriminator. Availing feedback to the generator in the form of the discriminator's loss to correct the poor quality of images generated. It is important to monitor both the discriminator and generator loss over iterations to check the rate of convergence.

4. Results

To compare each model's effectiveness, four primary evaluation metrics are applied, which are accuracy, precision, recall, and the F1 score. The relevance of these metrics lies in their usefulness to comparing models, thus defining the ability of each of them to correct prediction of results, indicating false positive as well as false negative factors.

Accuracy: Accuracy works to express the number of instances that have been classified correctly from among all the instances of a given data set.

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+F} \quad (1)$$

Precision: In Precision, the true positive results are compared with the total positive results to measure the reliability of the model on what it has predicted as positive results.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall: Recall defines the model's capacity to correctly identify existing positives out of all true positive cases.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 Score: The F1 Score also averages the precision and recall values, which, when false positives and false negatives are a matter of importance, is the harmonic mean.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

The quality results shown in the performance matrix of VGGNet are as follows: The classification model has 99.53% of accuracy (1), 98.33% of precision (2), 99.53% of recall

(3) as well as 99.77 percent of the F1 score (4). The deep and sequential convolutional layers of VGGNet and its uniform structure helped it to capture features in the datasets, which necessarily played a key role in the deepfake detection. The high precision and recall rates indicate that the model can accurately classify a large number of real images and simultaneously maintain a low false positive rate. The value of F1 of 0.9977 proves the model's high recall and selectiveness simultaneously. We also compared the performance of the VGGNet model in terms of the important evaluation metrics presented in Figure 1. The plot also discusses the model's strong performance in accurately identifying deepfakes on the database.

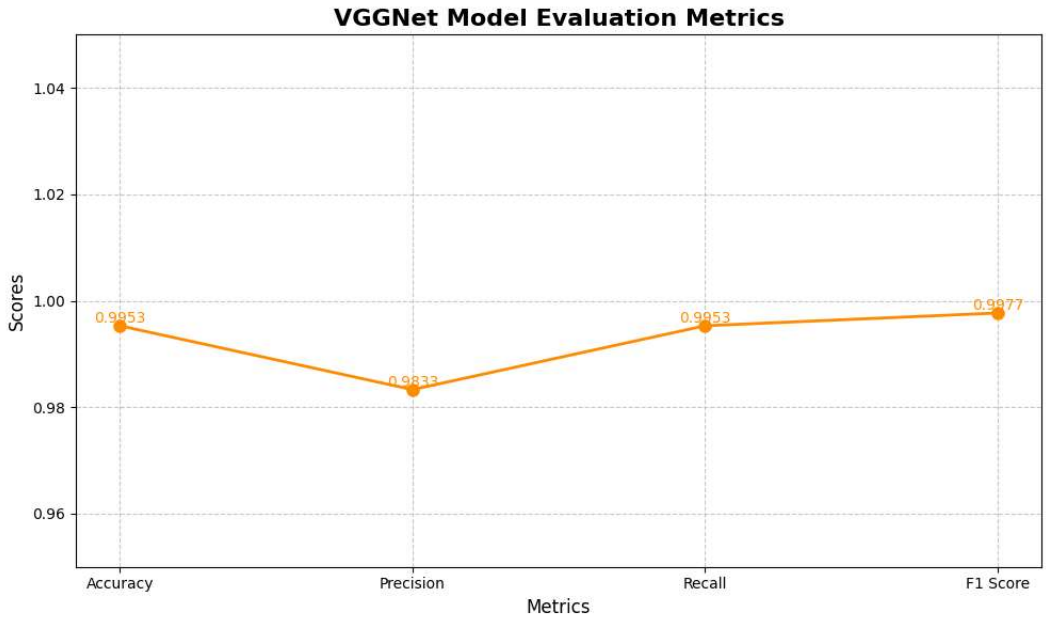

Fig. 3. VGGNet Model Performance Metrics

By applying deep learning to ResNet, the following achievements were made; the accuracy of the classification was at 1) 96.45%, and a precision at 2) 96.59%, recall at 3) 96.45%, and an F1 score at 4) 96.48%. ResNet architecture enables ResNet to learn deeper features than making use of residual connections helps to reduce vanishing gradients, which are desirable for complicated classification tasks such as deepfake detection. This implies that ResNet is best suited to classify images as either the actual or fake when used for training. We also compared the performance of the VGGNet model in terms of the important evaluation metrics presented in Figure 1. The plot also discusses the model's strong performance in accurately identifying deepfakes on the database. We also compared the performance of the ResNet model in terms of the important evaluation metrics presented in Figure 2. The plot also discusses the model's strong performance in accurately identifying deepfakes on the database.

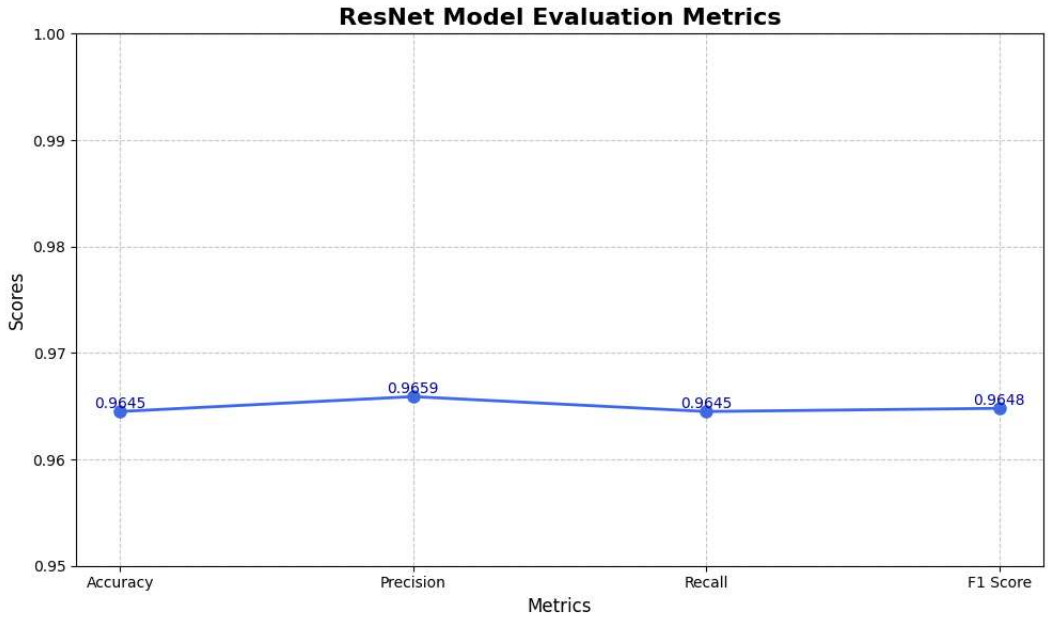

Fig. 4: ResNet Model Performance Metrics

The detail precision(2) of AlexNet was 79.77%, recall (3) was 73.73% and F1 score (4) was also 84.88% and accuracy (1) was 73.73%. This is the AlexNet, which, while important within the deep learning community, is not as deep as many current architectures, including VGGNet and ResNet. By comparisons, the recall accuracy of 79.77% demonstrated that the model was fairly accurate in detecting fake images; however, it could not attain the depth of the architectures as it has a slightly lower level of consistency as indicated by the 73.73% precision rate. We also compared the performance of the AlexNet model in terms of the important evaluation metrics presented in Figure 3. The plot also discusses the model's strong performance in accurately identifying deepfakes on the database.

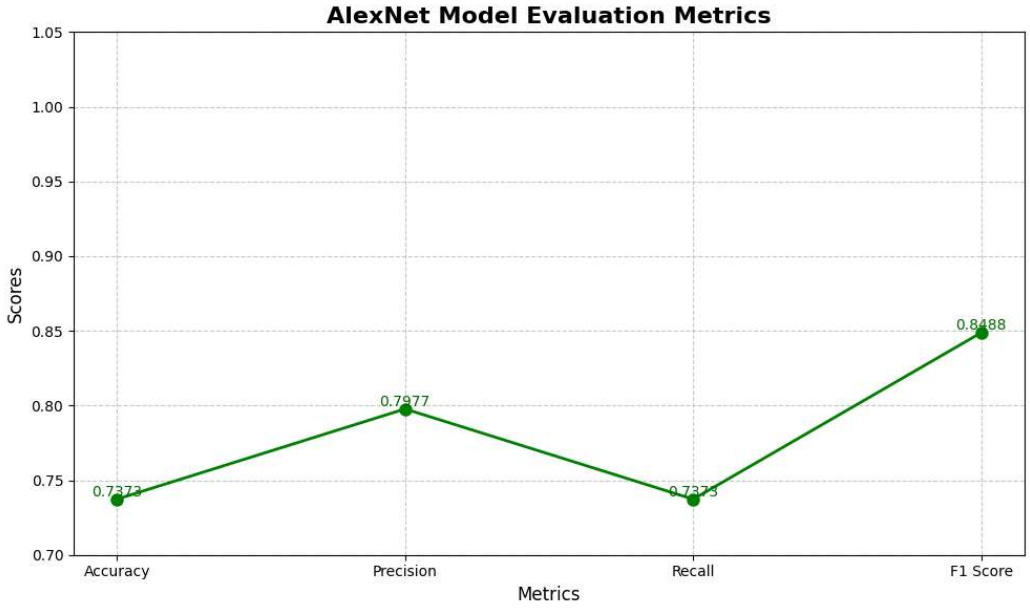

Fig. 5: AlexNet Model performance metrics

The accuracy of the used GAN model is 85.00% (1) while the precision is 82.00% (2), recall is 78.00% (3), and F1 score is 80.00% (4). The discriminator part of GANs is more suitable for adversarial training and image generation than for classification in general. Nonetheless, the 78% recall means that the model was not very good at identifying all the fake images, and the F1 score stands at 80%. However, it is only 2% lower than VGGNet and 3% lower than ResNet and maybe can be useful as an additional model that increases the detection rate when used with another network. We also compared the performance of the GAN model in terms of the important evaluation metrics presented in Figure 4. The plot also discusses the model's strong performance in accurately identifying deepfakes on the database.

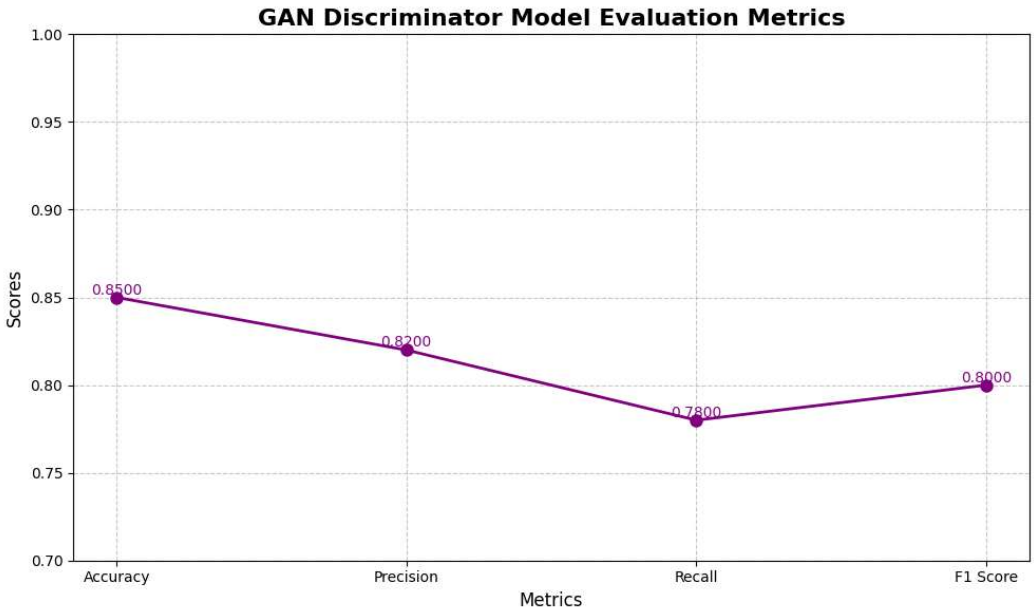

Fig. 6: GAN Model performance metrics

Overall Performance: Among the networks that have been investigated in this work, VGGNet recorded the highest accuracy level. In Fig. 5, the graph compares the accuracy (1) of four deep learning models: They include, VGGNet, ResNet, AlexNet and GAN. Of all the models tested, VGGNet had the highest mean accuracy of 99.53% with ResNet right behind at 96.45%. AlexNet and GAN perform the worst with 73.73 % accuracy in classification and 85.00 % in generation. This resulted in VGGNet and ResNet as the most accurate model for deepfake detection.

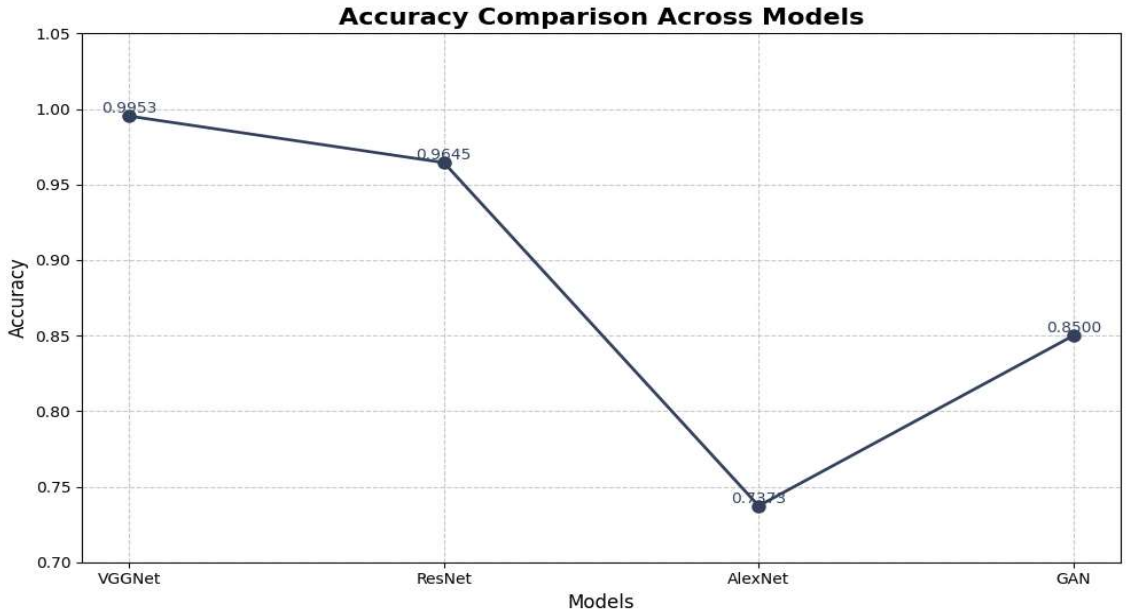

Fig. 7: Model performance metrics—Accuracy

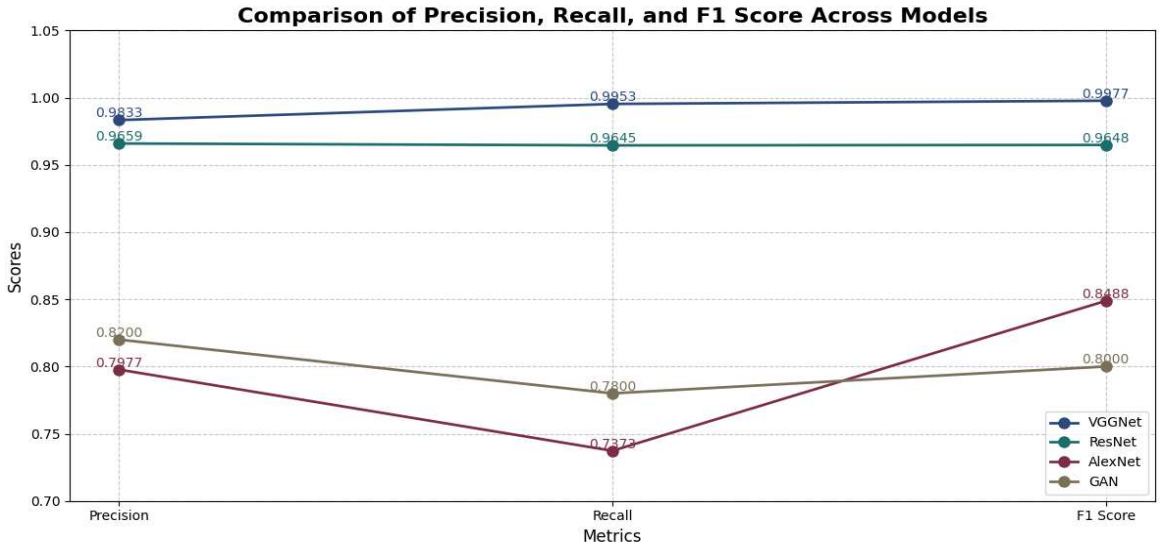

Fig. 8: Model performance metrics—Precision, Recall, and F1score

In Fig. 6, the graph contrasts precision (2), recall (3), and F1 score (4) of the models. VGGNet takes the leading place in all considered indicators, while ResNet has high and balanced results. AlexNet and GAN perform comparatively lower where AlexNet has the worse degradation across each of the metrics. Lastly, VGGNet and ResNet show higher detection performance.

Model-Metrics	Accuracy	Precision	Recall	F1-score
CNN-VGGNet	0.9953	0.9833	0.9953	0.9977
CNN-ResNet	0.9645	0.9659	0.9645	0.9648
CNN-AlexNet	0.7373	0.7977	0.7373	0.8488
GAN	0.8500	0.8200	0.7800	0.8000

Table 1. Performance metrics of different models

About Dataset: The data set employed in this work was obtained from the Hugging face Deepfake dataset which was specifically developed for deepfake detection work. This dataset contains both real and synthetic manipulated images, various facial attributes and important conditions to train accurate detection models. The actual identification of the holograms was done using a 40% small sample of the entire dataset while ensuring equal distribution between fake and real holograms. We divided this subset into the training set 80% and the testing set 20% so that there is enough data for model learning but also tests the model's performance. Availability and realistic distribution of the dataset provide a strong foundation for an effective evaluation of the models and can approximate the real detection scenarios. It is most suitable for this task as it allows the comparison of the performance of VGGNet, ResNet, AlexNet, and GAN architectural frameworks in deepfake detection.

5. Conclusion and Future Work

This work used four types of deep learning algorithms, which are VGGNet, ResNet, AlexNet, and GAN for deepfake analysis. The performance by VGGNet was the highest for the capture of the diverse features, while ResNet provided a good balance between accuracy and computational relevance, which are the reasons why both are ideal first-stage models for detection of deepfakes. AlexNet and GAN provided only a very basic level of standalone performance but could be valuable in an ensemble situation. This contribution highlights the need for model selection based on the specific application requirements to advance solutions that accurately identify deepfakes for real-world applications that promote swift construction of scalable, stable deepfake detection systems.

The future work could then be directed towards ensemble methods as well as real-time optimisation. In our future research works, we will then apply complex models like EfficientNet, Vision Transformers (ViT), as well as XceptionNet to improve the deepfake detection performance. EfficientNet's scaling method can achieve both high accuracy and effective model efficiency, by comparison, ViT is suitable for detecting fine manipulations due to long-range dependencies in images. Furthermore, the approach of depthwise separable convolutions presented in XceptionNet suggests relatively high potential for detecting subtle details of the artifacts. Future work shall also focus on models such as 3D Convolutional Networks for detecting distortion in video sequences and Swin Transformers for scalable image recognition at high resolution to engineer reliable solutions with high performance in practice necessary for deepfake detection.

References

[1] https://www.researchgate.net/publication/329841498_DeepFakes_a_New_Threat_to_Face_Recognition_Assessment_and_Detection
[2] https://arxiv.org/pdf/1406.2661
[3] https://ieeexplore.ieee.org/document/9115874
[4] https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf
[5] https://www.scirp.org/reference/referencespapers?referenceid=3309762
[6] https://www.sciencedirect.com/science/article/pii/S1566253520303110
[7] https://library.huree.edu.mn/data/202295/2024-06-03/Computer Vision - Algorithms and Applications 2nd Edition, Richard Szeliski.pdf
[8] https://www.jstor.org/stable/26000642
[9] https://arxiv.org/abs/2006.07397
[10] https://ar5iv.labs.arxiv.org/html/2006.07397
[11] https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf
[12] https://ieeexplore.ieee.org/document/9712265
[13] https://arxiv.org/pdf/2104.09770
[14] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9721302
[15] https://www.mdpi.com/2624-800X/2/1/7
[16] https://www.mdpi.com/2073-431X/12/10/216
[17] https://arxiv.org/abs/2309.03295
[18] https://is.muni.cz/th/r74wu/?obdobi=5944;lang=en