Deepfake Detection: A Comparative Analysis of Deep Learning Algorithms

Prashanthi Thota
Department of CSE
SRM University-AP
prashanthi_thota@srmap.edu.in

Sashank Sakkurthi
Department of CSE
SRM University-AP
sashankchowdary_s@srmap.edu.in

Lakshmi Nikhitha Dodda
Department of CSE
SRM University-AP
lakshminikhitha_d@srmap.edu.in

KPVM Karthik
Department of CSE
SRM University-AP
manikarthik_kakumanu@srmap.edu.in

Srilatha Tokala
Department of CSE
SRM University-AP
srilatha_tokala@srmap.edu.in

Murali Krishna Enduri
Department of CSE
SRMUniversity-AP
muralikrishna.e@srmap.edu.in

Abstract: **Deepfake has become instrumental in arousing several ethical and security issues, therefore demanding careful detection techniques. This study investigates deepfake detection through a comparative analysis of four advanced deep learning algorithms: three different types of convolutional neural networks, which are CNN-VGGNet, CNN-ResNet, CNN-AlexNet, and GANs (Generative Adversarial Networks). Leveraging the unique architectures and feature extraction capabilities of these algorithms, this study evaluates their efficacy in identifying deepfakes using standardized evaluation metrics: What we have is accuracy, precision, recall, and F1 score. The effectiveness of each algorithm is discussed in detail, and while CNN-based models are discussed for their high-dimensional feature extraction, GANs are discussed for adversarial training. In benchmarking these methods, this study offers the much-needed assessment of the characteristics and performance of each model where it counts in real detection scenarios.**

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

I. Introduction

Hyper real manipulative media exploited by deepfake tech has surged a prodigious machine learning advancement and a noteworthy social menace. They are employed in sensitive areas such as fake news, identity thefts, and cyber espionage; which are both ethically and security dangerously ailing society [1]. The basics of this technology are based on deep learning which holds a big potential in differentiating fake media from the original content [2]. However, Convolutional neural networks (CNNs) have proved to provide high effectiveness for visual recognition tasks hence useful for faking images and videos. During the hierarchical feature extraction, CNN implicitly preserves important features that are necessary for the spotting of fine details of artifacts in deepfake images [3]. For instance, VGGNet uses a deep and uniform CNN model that improves its vision of details and identifies modified parts of deepfake material [4]. ResNet builds on this structure by providing residual connections for fixing some problems, such as vanishing gradients and being able to detect high levels of networks [5]. Further, the generative adversarial networks (GANs) present a special adversarial setup introduced for synthesizing synthetic media, which is now essential for deepfake detection. Although GAN based approaches minimize false positives, their adversarial training framework can be slow and is not suitable for real time use. This research aims to assess and compare the feasibility of using the following models in the area of deepfake detection: CNN-VGGNet, CNN-ResNet, CNN-AlexNet, and GAN while identifying benefits and drawbacks connected with each of them. 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 artefacts inherent in deepfakes.

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 characterised by the possibility of correspondingly high accuracy while minimising the risks of overfitting. 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 because it does not take long to train and recognises images quite effectively, so it is still quite a good option for the deepfake detection.

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 minimise small discrepancies undetectable by other CNN-based models.

In this paper, we have compared four famous deep learning models which we have trained on the images of faces and they are VGGNet, ResNet, AlexNet and GAN for deepfake detection. By assessing these models with fundamental parameters as usually accuracy, precision, recall, F1 score, the work shows the advantages of the particular architectures and reveals the potential drawbacks of them. Furthermore, the work offers real-world applications on how to choose models capable of being used for application.

II. Related Work

Tolosana et al. [6] reviewed various methodologies about face manipulation and deep fake detection focusing on the CNN and GAN approaches for handling face forgery. Szeliski [7] has done a lot to the advancement of computer vision to produce algorithms and applications that can be used for recognizing deepfake manipulations through the use of advanced vision. This research [8] on digital image forensics described the techniques that can be used to detect fraud images and the new techniques to defend against fake images. Dolhansky et al. [9] proposed the DFDC dataset pointing to the need to diversify datasets in order to ensure better generality of the models and, consequently, the reliability of the detection. The Celeb-DF was proposed by Li et al. [10]-it creates more challenges to improve deepfake technologies for forensic applications by including complicated manipulations for the detectors. Malik et al. [11] reviewed deepfake detection for facial images and videos currently and pointed out that there are some key problems with detecting deepfake videos at high real-world accuracy.

III. Materials and 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 [12]. 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 [13]. Further, the given section describes the method used to implement each model for further details regarding the procedure.

A. CNN-VGGNet: On the dataset side, load the dataset and include both real and fake images [14]. We resize these images such that they fit VGGNet's input layer at 224x224 in size. Dividing pixel values by 255 normalises 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. The base model initialises with VGG16, already pre-trained with ImageNet [15]. 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. Regularisation by dropout and batch normalisation layers is used to regularise 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.

B. CNN-ResNet: Clean the dataset to ensure they meet ResNet specifications by resizing the images to 224*224 pixels. Standardise 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 [16]. ResNet50 is selected, and the pre-trained model is ResNet50, which is also initialised 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 [17]. Add additional fully connected layers: This is followed by the Global Average Pooling Layer, then followed by the dense layers incorporating batch normalisation and dropout. Select categorical cross entropy for regression; optimize as Adam at a rate of 0.0001.

C. CNN-AlexNet: All images should be resized to fit 227 such that they meet the input size for the AlexNet model. Within the pixel values, standardise them, and then put through data augmentation as was done with other CNN models above. In order to implement AlexNet, we start with the characteristic of having in the network successive layers of convolution and pooling [18]. 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.

D. GAN: Resize the images to fit a 64 x 64 input frame that is the architecture of the GAN [19]. Then, resize images to 64x64 and normalise the pixel range to [-1, 1], which is the

output range of the generator's tanh activation function. 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 minimise 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 [20]. It is important to monitor both the discriminator and generator loss over iterations.

IV. DATASET

The dataset employed in this work is a well-selected list of data taken from the Hugging Face Deepfake Dataset. The images are organised into two classes: Real- It stands for actual and original faces and Fake- Stands for fake images created with deep fake technology. The dataset comprises various aspects of facial properties such as illumination changes, facial expression, and orientation, which are central for building stable detection models. To prepare the data for this study: These images were reduced to 224 by 224 pixels since most of the neural networks, including VGGNet, ResNet, and AlexNet, operate with this size. In the case of the GAN-based model, the images were dimensioned 64 x 64 pixels in order to meet the requirements of this model. Dividing dataset in to 80% training and 20% testing data was done so as to have enough data to learn from while being as stringent as possible in the testing process. Normalisation was done to ensure that pixel values fell within the range 0 to 1 to enhance the models computational efficiency. Data augmentations including rotation, flipping alongside zoom was incorporated to prevent overfitting as well as increase the model's generalisation proficiency.

Fig. 1: Example Images of Fake and Real Faces from the Dataset

V. RESULT ANALYSIS

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: As shown in Eq. (1), Accuracy works to express the number of instances that have been classified correctly from among all the instances of a given dataset.

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

Precision: As shown in Eq. (2), 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: As shown in Eq. (3), 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: As shown in Eq. (4), 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.

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

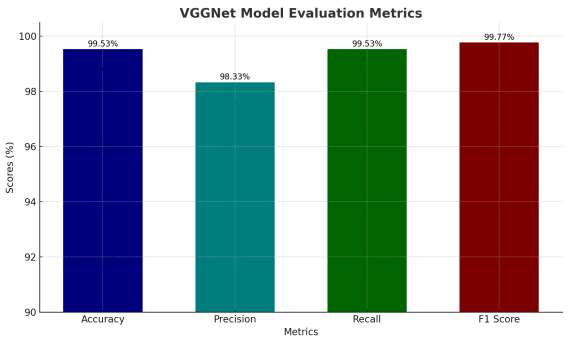

Fig. 2: VGGNet Model Performance Metrics

The quality results shown in the performance matrix of VGGNet are as follows: The classification model has 99.53% of accuracy, 98.33% of precision, 99.53% of recall as well as 99.77 percent of the F1 score. 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 Fig.2.

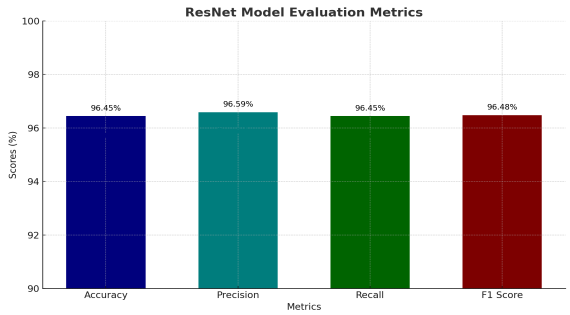

Fig. 3: ResNet Model Performance Metrics

By applying deep learning to ResNet, the following achievements were made; the accuracy of the classification was at 96.45%, and a precision at 96.59%, recall at 96.45%, and an F1 score at 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 ResNet model in terms of the important evaluation metrics presented in Fig.3.

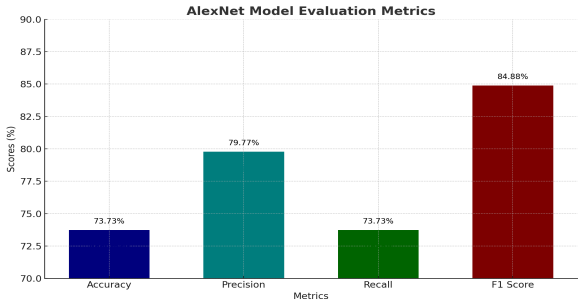

Fig. 4: AlexNet Model performance metrics

The detail precision of AlexNet was 79.77%, recall was 73.73% and F1 score was also 84.88% and accuracy 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; 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 Fig.4.

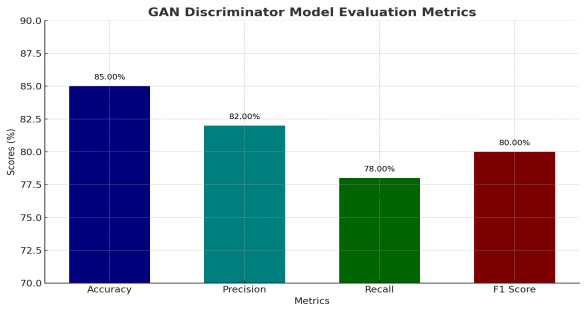

Fig. 5: GAN Model performance metrics

The accuracy of the used GAN model is 85.00% while the precision is 82.00%, recall is 78.00%, and F1 score is 80.00%. 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 Fig.5. The plot also discusses the model's strong performance in accurately identifying deepfakes on the database.

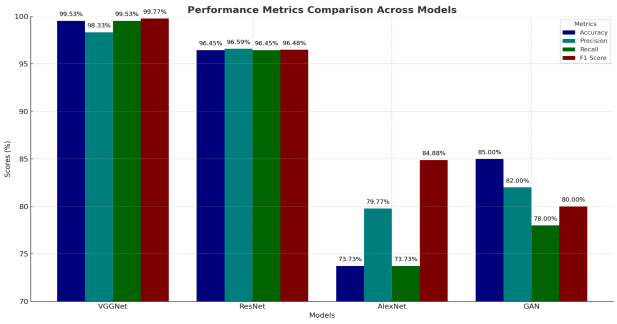

Fig. 6: Model performance metrics

Overall Performance: The bar graph in Fig. 6 presents a comprehensive comparison of the performance metrics: Accuracy, Precision, Recall, and F1 Score across the four deep learning models: The models include VGGNet, ResNet, AlexNet and GAN. The experimental results of the VGGNet model have the highest level of accuracy, close to ideal, consisting of 99.53% Accuracy and 99.77% F1 Score. Precision of ResNet is found to be 96.59%, while recall is 96.45% marking a balanced and quite stable network which does not yield high false positives and false negatives. AlexNet performs comparatively lower in most of the parameters; it has low but equal to the lowest Recall of 73.73% showing its poor ability for true positive detection. Still, it gives a relatively higher F1 Score of 84.88% because of reasonable Precision and Recall. As for

the result, GAN has relatively good results with Accuracy of 85.00% and Precision of 82.00% slightly higher than AlexNet. Still, recall of 78.00%, and F1 Score 80.00% show that it might not identify more complex patterns as the other models do. The overall performance across different models is tabulated in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
VGGNet	99.53%	98.33%	99.53%	99.77%
ResNet	96.45%	96.59%	96.45%	96.48%
AlexNet	73.73%	79.77%	73.73%	84.88%
GAN	85.00%	82.00%	78.00%	80.00%

Table.1 Comparative Analysis of Performance Metrics Across Models

VI. 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]. Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

[2]. Goodfellow, P. A., Mirza, X., Warde-Farley, O., & Courville, B. (2014). Goodfellow I. Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., Generative adversarial nets, Advances in Neural Information Processing Systems, 2672-2680.

[3]. Verdoliva, L. (2020). Media forensics and deepfakes: an overview. IEEE journal of selected topics in signal processing, 14(5), 910-932.

[4]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

[5]. Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[6]. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

[7]. Szeliski, R. (2022). Computer vision: algorithms and applications. Springer Nature.

[8]. Farid, H. (2008). Digital image forensics. Scientific American, 298(6), 66-71.

[9]. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.

[10]. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).

[11]. Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection for human face images and videos: A survey. *Ieee Access*, *10*, 18757-18775.

[12]. Kumar, M., & Sharma, H. K. (2023). A GAN-based model of deepfake detection in social media. *Procedia Computer Science*, *218*, 2153-2162.

[13]. Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE access, 10, 25494-25513.

[14]. Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy, 2(1), 89-106.

[15]. Chaudhary, S., Saifi, R., Chauhan, N., & Agarwal, R. (2021, December). A comparative analysis of deep fake techniques. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 300-303). IEEE.

[16]. Sontakke, N., Utekar, S., Rastogi, S., & Sonawane, S. (2023). Comparative Analysis of Deep-Fake Algorithms. arXiv preprint arXiv:2309.03295.

[17]. SVOBODA, J. AI Deep Fake Perception: A Critical Comparative Research And Study Design.

[18]. Chang, X., Wu, J., Yang, T., & Feng, G. (2020, July). Deepfake face image detection based on improved VGG convolutional neural network. In *2020 39th Chinese control conference (CCC)* (pp. 7252-7256). IEEE.

[19]. Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. *The Visual Computer*, *39*(12), 6323-6344.

[20]. Rafique, R., Nawaz, M., Kibriya, H., & Masood, M. (2021, November). Deepfake detection using error level analysis and deep learning. In *2021 4th International Conference on Computing & Information Sciences (ICCIS)* (pp. 1-4). IEEE.