2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-6263-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/Confluence56041.2023.10048888

A Comparative Study: Deepfake Detection Using Deep-learning

Nishika Khatri
Amity School of Engineering and Technology, Amity University
Uttar Pradesh
nishikakhatri2611@gmail.com

Varun Borar
Amity School of Engineering and Technology, Amity University
Uttar Pradesh
varunborar@gmail.com

Rakesh Garg
Amity School of Engineering and Technology, Amity University
Uttar Pradesh
rkgarg06@gmail.com

***Abstract*—In recent decades, we have seen significant advancement in fields like Artificial Intelligence, Machine Learning, and Deep Learning, resulting in the developing of new technologies such as deepfake. Deepfakes are a form of digital media that replaces one identity's likeness with another or creates a synthetic personality: in the form of high-quality realistic fake video, image, or audio. Deepfakes can be helpful in education, art, activism, and self-expression; however, some subjects can use deepfakes to harm the portrayal of people, create pornographic content, and spread misleading information. High-quality deepfakes are easy to build but incredibly difficult to detect, creating a need to explore technologies which can be helpful in deepfake detection. Therefore, we present a comparative study of deep-learning models that can benefit deepfake detection. We have explored four deep-learning models, namely, VGG16, MobileNetV2, XceptionNet, and InceptionV3 and trained these models on the FaceForensics++ dataset. Finally, we evaluate the performance of these models for deepfake detection and conclude the study with our observations and future scope for improvement in this field.**

Keywords—deepfake detection, computer vision, deep learning, image classification, convolutional neural networks

I. Introduction

Deepfakes are a form of digital media used to replace one person's likeness with another or forge a synthetic face, voice, or expression. The advancements in this field are mainly driven due to the ascent in artificial intelligence, machine learning and deep learning [1]. Generative Adversarial Networks (GANs), introduced in 2014, have numerous applications in computer vision; they are extensively used in deepfake creation. They usually consist of two competing neural networks, one which creates the forged media and one which detects the forgery to obtain realistic-looking images.

Deepfakes can be helpful in education, art, activism and self-expression; however, some subjects can use deepfakes to harm the image of people, create pornographic content, spread misleading information, and spread fear or disgust in people [2]. Over the past few years, giant steps in the field of photography and cinematography have refined facial manipulation methods [3]. However, unfortunately, it has become effortless for anyone to create high-quality deepfakes that are hard to distinguish for the human eye and sometimes even the computer. Thus, it becomes challenging to verify the authenticity of the media, which poses severe problems for social media platforms, banks, and academic institutions.

With increasing interest in the domain, numerous research on deepfake detection has come forward. Datasets such as CELEB-DF [4], FaceForensics++, and Wilddeepfake [5], amongst others, have surfaced, which can be utilized for the problem. Though several advancements have been achieved, many critical issues must be resolved for the existing deepfake detection methods. Furthermore, with the evolving quality of deepfakes, some traditional methods are no longer helpful [6]. This study compares the various methods for deepfake detection on a common dataset and set of parameters.

The Section II introduces to readers about the problem and scrutinizes the work done in the field. Then, it elucidates the classification of deepfake detection methods into statistical models, machine learning and deep learning. Section III discloses the research questions and objectives of the study along with the dataset used, data pre-processing techniques, metrics used and the experimental setup for the study. The Section IV presents readers with a comparative analysis of the various deep learning models trained and tested on the FaceForensics++ dataset. Finally in Section V and Section VI, we summarise the research findings and outline the domain's future scope.

II. Literature Review

Deepfake threaten the privacy, integrity and reliability of information made available in media. The domain of deepfake has garnered much attention, and multiple deepfake detection techniques have evolved in recent years. M. S. Rana et al. [7] have classified deepfake detection methods into statistical, machine learning, and deep learning models.

A. Statistical Models for Deepfake Detection

Statistical models include algorithms like Expectation Maximization (EM) [8], Kullback-Leibler Divergence (KLD), Total Variational Distance (TVD) and Jensen-Shannon Divergence (JSD) [9]. EM can extract convolutional traces directly from the images [8]. However, statistical models limit the accuracy of deepfake detection on the relative resolution of images and the accuracy of GAN, i.e., a higher resolution image requires equally low accuracy of the GAN that is employed to generate the synthetic image [10]. However, over the years, the accuracy of GANs has undeniably improved to a point where they can create realistic-looking, high-resolution images from scratch, which makes it challenging to employ statistical models for highly accurate deepfake detection.

B. Machine Learning for Deepfake Detection

Machine Learning models include Support Vector Machine (SVM) [11], Logistic Regression (LR), Multilayer Perceptron (MLP), K-Means Clustering (K-MN), Multiple Instance Learning (MIL), and Naive Bayes (NB), among others. It creates a feature vector using a feature selection algorithm, and then the vector is fed as an input to train the classifier to predict whether the media is manipulated. M. S.

Rana, B. Murali, and A. H. Sung [12] explain that feature extraction and selection are significant problems in machine learning models. Machine learning models provide better understandability and interpretability along with reduced training time, but for enhanced model performance, it is necessary to identify and discriminate relevant features accurately. This problem can be subdued with deep-learning models.

C. Deep Learning for Deepfake Detection

Deep learning models are widely used for applying deepfake detection due to their feature extraction and selection mechanism ability. Deep Learning models can be further categorized in the domain of Convolutional Neural Networks (CNN) (e.g., MobileNetV2, XceptionNet, VGG etc.), Recurrent Neural Networks (RNN) [13], Region-based Convolutional Neural Networks (RCNN), Hierarchical Memory Network (HMN) [14] and Multi-task Cascaded CNN (MTCNN). CNN, RCNN, and MTCNN models are extensively used for feature detection and extraction. RNN can capture temporal inconsistencies between the frames [15]. HMN can be used to mitigate the problem of generalization; it preserves the characteristics of previously processed faces in the memory [14]. In this study, we mainly focus on CNN models, including, MobileNetV2 [16], InceptionV3 [17], VGG16 [18] and XceptionNet [19], for the problem of deepfake detection.

III. Methodology

A. Research Objectives

Through our study, we aim

1. Identify and group available deepfake datasets and select one for further research.
2. To identify the existing methods used for deepfake detection.
3. To determine the accuracy and other parameters and prepare a comparative study by testing models on the selected datasets.

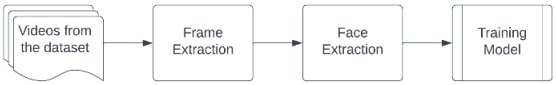

Fig. 1. Training process

B. Dataset

We have utilized the publicly available FaceForensics++ Dataset [20] with a compression factor of 23. The dataset contains 4000 deepfake videos with an average of 470 frames per video and 1000 original sequences. Deepfake videos are generated from original sequences using four different GAN, and thus, can be classified into four types based on the GAN used for their creation, namely,

(a) Deepfakes

(b) Face2Face

(c) FaceSwap

(d) NeuralTextures

(e) Original

Fig. 2. Sample images from the dataset

1. Deepfakes
2. Face2Face
3. FaceSwap
4. NeuralTextures

C. Data Pre-Processing

On analysis, we could determine that the consecutive frames in each video are redundant; thus, we have extracted every 28th frame from each video. Following the process, we extracted faces with a 30% margin from the frames to create the final dataset for evaluating the models. TABLE I. shows the total size of the processed dataset.

TABLE I. No. Of Frames For Each Dataset

Dataset	No. of Frames Extracted with Faces
Deepfakes	18580
Face2Face	18636
FaceSwap	14954
NeuralTextures	14953
Original Sequences	18635

As the number of frames for the FaceSwap and NeuralTextures dataset is significantly less than the original sequences, while training the model, we have used an under-sampling procedure to randomly select an equal number of frames.

D. Metrics Used For Comparative Analysis

Many different models can be used for the problem of deepfake detection. Even within a model, several parameters, such as activation functions, loss functions, optimizers, and learning rate, can affect the results. Therefore, for performing a comparative study, we must evaluate all the models with similar parameters, and the results should be compared against a standard set of metrics.

While many metrics exist for evaluating the performance of deep learning models, none alone offer a complete analysis of the model's performance. Hence, we use the following set of metrics to compare the various aspects of each of the models:

1) Accuracy

Accuracy measures how close the predicted values by the model are to the true values. Since this is a classification problem, the accuracy is calculated simply as the ratio of the images classified correctly by the model to the total number of images. However, since it is a simplistic metric (it does not account for the other aspects of the model), additional measures are necessary to characterize and evaluate a model adequately [21].

$$Accuracy = \frac{No.of\ Correct\ Predictions}{Total\ No.of\ Prections} - (1)$$

2) Precision

Precision attempts to measure how many identifications are correct. It is calculated as,

$$Precision = \frac{No.of\ True\ Positives}{No.True\ Positives + No.False\ Positives} - (2)$$

A true positive is defined as an image that is positively classified into a particular class by the model that belongs to that class. On the other hand, a false positive is defined as an image positively classified into a particular class by the model which does not belong to that class [21].

Precision can be used to determine the relevancy of the classification results produced by the model. A higher precision means that the results are more often relevant than not, while a lower precision signifies that the model may categorize the images wrong more often.

3) Recall

Recall attempts to measure what proportion of the total number of actual positives are identified correctly for a particular class. Mathematically,

$$Recall = \frac{No.True\ Positives}{No.True\ Positives + No.False\ Negatives} - (3)$$

A false negative is defined as an image negatively classified into a particular class by the model which does not belong to that class.

A high recall signifies that the model accurately captures the image class, i.e., the image is classified accurately. This metric counteracts precision and ensures to provide a complete picture of the model [21].

4) F_1-Score

As stated, precision and recall are counteractive measures to each other, i.e., an attempt to increase one metric tends to decrease the other. However, they are both critical to analyze the model.

To provide cumulative results of both metrics, F1-score may be utilized [22]. F1-score is the harmonic mean of precision and recall. It can be used independently to measure the model's performance. A high F1-Score signifies a better performance, i.e., a higher recall and precision. A low F1-score means that either or both metrics are low. It can be defined as follows,

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} - (4)$$

Further, for comparing the performance of models across the different datasets and providing cumulative results, we have also used basic statistical metrics such as mean, median and variance.

E. Experimental Setup

1) Hardware:

We utilized a cloud-based system with 150 GB of persistent storage, 8 vCPU, 45 GiB RAM, and 16 GiB GPU (NVIDIA RTX A4000).

2) Software:

All the experiments were done on a Linux-based instance, with Python 3.9, the latest versions of Tensorflow (2.9) and Keras API (2.10).

IV. Result and Discussion

Initially, we used the concept of transfer learning with imagenet weights for feature extraction, but due to the difference in the dataset, output classes and low accuracy of the trained models, we switched to training the entire architecture from scratch.

We have set the input image size for each model to 224x224x3 (channel last configuration) and replaced the default top layer (used for imagenet classification into 1000 classes) with a fully connected layer that results in an output of 2 classes.

TABLE II. Cumulative Accuracy And Variance

Dataset	Mean	Median	Variance
MobileNetV2	87.98%	89.65%	100.87
InceptionV3	95.54%	96.98%	14.88
VGG16	95.14%	97.19%	26.56
XceptionNet	95.84%	97.55%	12.43

A. MobileNetV2

It is a lightweight convolutional neural network architecture that improves the state-of-the-art performance of mobile models on multiple tasks and acts as a benchmark across a spectrum of different model sizes [16]. It consists of two main components: inverted residual block and bottleneck residual block. There are 53 convolutional layers, which fall under 1x1 convolutional, and 3x3 depthwise convolutional layers with one AvgPool layer. MobileNetV2 follows the narrow-wide-narrow approach, which means that each block has three different layers, i.e., 1x1 Convolutional with Relu6, 3x3 Depthwise convolution, and lastly 1x1 Convolutional layer without any linearity. Each convolutional layer in MobileNetV2 consists of input height and width, input channel, kernel, stride, padding, and output.

TABLE III. shows the testing result of the MobileNetV2 when tested on different datasets. The results show that the model converged very well on the Deepfakes and Face2Face datasets but failed to provide a similar outcome for Facswap and NeuralTextures datasets.

B. InceptionV3

It is an advance and optimized version of InceptionV1, which Google developed in 2015. InceptionV3 uses optimization techniques for better model performance, such as factorization into smaller convolutions, spatial factorization into asymmetric convolutions, utility of auxiliary classifiers, and efficient grid size reduction [17]. The InceptionV3 model is made up of 42 layers. The architecture of InceptionV3 is more profound than InceptionV1 and InceptionV2; it provides higher efficiency and less error rate and is computationally less expensive.

TABLE IV. shows the testing results of InceptionV3. Compared to MobileNetV2, the model converged better on the Deepfakes and Face2Face datasets and performed far better on the other two. However, the comparative performance of the model produces significant variance when tested across different datasets.

C. VGG16

VGG16's architecture is based on a deep neural network, a variant of the VGG model. As the name suggests, it consists of 16 convolutional layers and has a uniform architecture [18]. It only contains convolution and pooling layers. It uses a 3x3 kernel size for the convolutional layer and a 2x2 size for the maxpool layer.

TABLE V. shows the testing results of VGG16. Similar to InceptionV3, it performed very well on Deepfakes, Face2Face and FaceSwap datasets but could not give similar results for the NeuralTextures dataset. Moreover, the variance for this model is significantly higher than the InceptionV3.

D. XceptionNet

It is an extreme version of the Inception deep learning model based on depth-wise separable convolutional layers. It has 36 convolutional layers, which form the base for feature extraction. It applies the filter on each depth map and then compresses the input space using 1x1 convolution [19]. XceptionNet differs from Inception models in order of operations, i.e., Inception performs 1x1 convolution first, whereas XceptionNet performs channel-wise spatial convolution and then performs 1x1 convolution. Another point of difference is that in Inception, both operations are followed by a ReLU non-linearity, while in XceptionNet, separable convolutions are implemented without non-linearities.

TABLE VI. shows the testing results of XceptionNet. While its performance was slightly less than InceptionV3 and VGG16 on the Deepfakes, Face2Face and FaceSwap datasets, it performed significantly well on the NeuralTextures dataset compared to others. Moreover, it has the lowest variance among the four models.

TABLE III. PERFORMANCE STATISTICS OF MODEL MOBILENETV2

Dataset	Accuracy	Loss	Precision-Fake	Precision-Real	Recall-Fake	Recall-Real	F1 Score-Fake	F1 Score-Real
Deepfakes	0.9730	0.1157	0.98	0.97	0.97	0.98	0.97	0.97
Face2Face	0.9467	0.1560	0.95	0.95	0.95	0.95	0.95	0.95
FaceSwap	0.8463	0.4604	0.87	0.83	0.82	0.87	0.84	0.85
Neural Textures	0.7534	0.5796	0.78	0.73	0.71	0.8	0.74	0.76

TABLE IV. PERFORMANCE STATISTICS OF MODEL INCEPTIONV3

Dataset	Accuracy	Loss	Precision-Fake	Precision-Real	Recall-Fake	Recall-Real	F1 Score-Fake	F1 Score-Real
Deepfakes	0.9805	0.0549	0.98	0.98	0.98	0.98	0.98	0.98
Face2Face	0.9822	0.0581	0.99	0.98	0.97	0.99	0.98	0.98
FaceSwap	0.9592	0.1248	0.96	0.96	0.96	0.95	0.96	0.96
Neural Textures	0.8997	0.3093	0.91	0.89	0.89	0.91	0.90	0.90

TABLE V. PERFORMANCE STATISTICS OF MODEL VGG16

Dataset	Accuracy	Loss	Precision-Fake	Precision-Real	Recall-Fake	Recall-Real	F1 Score-Fake	F1 Score-Real
Deepfakes	0.9823	0.0464	0.98	0.98	0.98	0.98	0.98	0.98
Face2Face	0.9860	0.0356	0.99	0.98	0.99	0.99	0.99	0.99
FaceSwap	0.9615	0.1188	0.97	0.95	0.95	0.97	0.96	0.96
Neural Textures	0.8758	0.3257	0.90	0.86	0.85	0.90	0.87	0.88

TABLE VI. PERFORMANCE STATISTICS OF MODEL XCEPTIONNET

Dataset	Accuracy	Loss	Precision-Fake	Precision-Real	Recall-Fake	Recall-Real	F1 Score-Fake	F1 Score-Real
Deepfakes	0.9772	0.0626	0.98	0.98	0.98	0.98	0.98	0.98
Face2Face	0.9747	0.0787	0.98	0.96	0.96	0.99	0.97	0.97
FaceSwap	0.9764	0.1070	0.98	0.97	0.97	0.98	0.98	0.98
Neural Textures	0.9056	0.4725	0.91	0.90	0.90	0.91	0.91	0.91

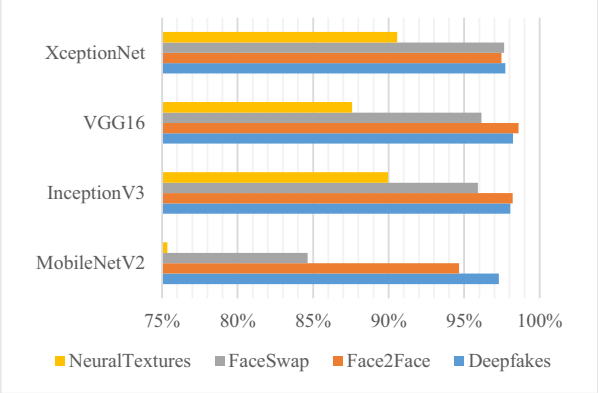

Fig. 3. Comparison of accuracy

V. Conclusion

In this study, we have performed a comparative analysis of the performance of four models - MobileNetV2, Face2Face, FaceSwap and XceptionNet – by training them on the publicly available FaceForensics++ dataset. Following this, we have presented the results of our study in the form of various parametric comparisons.

We have also shown that a model with the same training configuration and similar tuning parameters gives different results on different types of deepfakes images. The feature detection and accuracy of models are significantly higher on the Deepfakes dataset, followed by Face2Face, FaceSwap and NeuralTextures. The variance in accuracy is least for XceptionNet, followed by InceptionV3, VGG16 and MobileNetV2.

VI. Future Scope

We have shown that the performance of the models varies on different types of deepfakes. At the same time, it may be noted that a slight modification to the architecture and usage of data-augmentation techniques can significantly improve the performance and feature extraction of the model and decrease the variance amongst the different datasets. These objectives are left for future investigations.

VII. References

[1] S. Negi, M. Jayachandran, and S. Upadhyay, "Deep fake : An Understanding of Fake Images and Videos," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 183–189, 2021, doi: 10.32628/cseit217334.

[2] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[3] S. Pashine, S. Mandiya, P. Gupta, and R. Sheikh, "Deep Fake Detection: Survey of Facial Manipulation Detection Solutions," 2021, [Online]. Available: http://arxiv.org/abs/2106.12605.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3204–3213, 2020, doi: 10.1109/CVPR42600.2020.00327.

[5] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection," *MM 2020 - Proc. 28th ACM Int. Conf. Multimed.*, pp. 2382–2390, 2020, doi: 10.1145/3394171.3413769.

[6] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A Survey on Deepfake Video Detection," *IET Biometrics*, vol. 10, no. 6, pp. 607–624, 2021, doi: 10.1049/bme2.12031.

[7] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[8] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020, doi: 10.1109/ACCESS.2020.3023037.

[9] S. Agarwal and L. R. Varshney, "Limits of Deepfake Detection: A Robust Estimation Viewpoint," pp. 1–7, 2019, [Online]. Available: http://arxiv.org/abs/1905.03493.

[10] T. T. Nguyen *et al.*, "Deep learning for deepfakes creation and detection: A survey," *Comput. Vis. Image Underst.*, vol. 223, no. July, p. 103525, 2022, doi: 10.1016/j.cviu.2022.103525.

[11] H. Agarwal, A. Singh, and D. Rajeswari, "Deepfake Detection Using SVM," *Proc. 2nd Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2021*, pp. 1245–1249, 2021, doi: 10.1109/ICESC51422.2021.9532627.

[12] M. S. Rana, B. Murali, and A. H. Sung, "Deepfake Detection Using Machine Learning Algorithms," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2021, pp. 458–463.

[13] Y. Al-Dhabi and S. Zhang, "Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)," in *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, 2021, pp. 236–241, doi: 10.1109/CSAIEE54046.2021.9543264.

[14] T. Fernando, C. Fookes, S. Denman, and S. Sridharan, "Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks," vol. 14, no. 8, pp. 1–16, 2019, [Online]. Available: http://arxiv.org/abs/1911.07844.

[15] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," *Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill.*, 2019, doi: 10.1109/AVSS.2018.8639163.

[16] M. Sandler, A. Howard, M. Zhu, and A. Zhmoginov, "Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.pdf," pp. 4510–4520, 2018.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[20] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 1–11, 2019, doi: 10.1109/ICCV.2019.00009.

[21] Ž. Vujović, "Classification Model Evaluation Metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.

[22] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," *Lect. Notes Comput. Sci.*, vol. 3408, pp. 345–359, 2005, doi: 10.1007/978-3-540-31865-1_25.