Study on Deepfake Face Detection using Transfer Learning Approach

Jannatul Mawa
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

Md. Humayun Kabir
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

ABSTRACT
The emergence of deepfake technology has added a new dimension to digital media manipulation. The increasing prevalence of these manipulated visual contents poses significant threats to the authenticity and trustworthiness of digital media. In response to this growing threat, this research work investigates into an approach for detecting deepfake face images through the fusion of transfer learning and deep ensemble neural network techniques. This methodology adopts transfer learning and ensemble neural network techniques to improve the accuracy of detection and generalization capabilities of deepfake detection models. The research includes an extensive evaluation of the deep ensemble neural network on available challenging deepfake datasets. The effectiveness of the applied strategy is evaluated against currently existing techniques using performance indicators such as accuracy, precision, recall, and F1-scores. Finally, this paper presents a notable contribution to the area of deepfake detection through the development of a transfer learning-based deep ensemble neural network.

Keywords
digital media, manipulation, ensemble neural network, detection models, accuracy, generalization capabilities.

1. INTRODUCTION
The rise of deepfake technology has led to significant concerns regarding the authenticity and credibility of multimedia content. Deepfake images, which are fabricated using sophisticated artificial intelligence techniques, can convincingly alter real-world scenarios by superimposing the face of one individual onto another in a seamless manner [1]. Detecting and preventing the dissemination of deepfake images has become a crucial challenge in preserving the integrity of visual information on digital platforms. This paper presents an approach, namely the transfer learning-based deep ensemble neural network designed to enhance the effectiveness of deepfake face detection.

The aim of this research is to detect deepfake face images with high accuracy than other existing methods using efficient models. To evaluate how existing pre-trained models perform in detection of deepfake images is also investigated. An ensemble of multiple models is applied on the dataset to improve classification accuracy, and fine-tuning is applied to the pre-trained models to improve the detection accuracy. Finally, the research shows how all of the factors contributes to enhancing the overall accuracy of the ensemble classifier.

2. RELATED WORK
Ali Raza et al. introduced a novel deepfake predictor (DFP) methodology that combines the VGG16 architecture with a convolutional neural network architecture with a precision of 95% and an accuracy of 94% in the task of deepfake identification [2]. Younis E. Abdalla et al. proposes a transfer learning approach for training image forgery detection models using deep learning techniques for detecting image fraud with obtaining validation accuracy of 94.89% in the task of image modification detection [3]. Kumar et al. employed two methodologies for the detection of deepfake photos. This method involved the development of a customized convolutional neural network (CNN) using deep learning techniques. Another approach involved transfer learning leveraging the pre-trained models to enhance the detection of deepfake images [4]. Chang et al. incorporated image noise and augmentation to a VGG network resulting in a new network NA-VGG, which made a lot progress over other cutting edge fake image detectors [5]. Ensemble learning is a methodology which involves training numerous models on a common dataset and then their predictions are aggregated. The objective of ensemble learning is to enhance performance by combining multiple models than the performance of any single model [6]. Qureshi et al. made use of six different base-learners, their predictions, along with the metadata are inputted into an SVM classifier that acts as the meta classifier. Where each base-learner has low accuracy, the meta-classifier outperforms them [7]. Sasikala et al. suggested an autonomous plant disease diagnosis method based on deep ensemble neural networks (DENN). Transfer learning approach is employed to reuse previously trained models. DENN outscored state-of-the-art pre-trained models by aggregation of different models, achieving an accuracy of 100% [8]. Sharma et al. presented a model using transfer learning technique from previously trained deep models like VGG16 and ResNet50 [9]. The proposed model is evaluated using three standard datasets. With the ensemble model reaching accuracies of 98.79%, 75.79%, and 95.52% on the three datasets, respectively, the overall performance is significantly improved [9]. Shad et al. used a number of techniques to identify deepfake photos and do a comparative study. With 99% accuracy, VGGFace prevailed over all the other models being looked into [10]. In a comprehensive assessment of the literature, Rana et al. observed that among four different methods, deep learning-based approaches perform better than other methods in deepfake image detection, according to an evaluation of the performance of various methods with regard to different datasets [11].

3. BACKGROUND STUDY
The following section provides a brief description of different types of learning techniques and example models called base learners.

3.1 Machine Learning

Machine learning is a type of AI technique that allows machines to analyze and learn useful patterns from data and make decisions out of them [12]. In this method, intelligent algorithms are trained with data rather than being explicitly programmed. During image processing, the images in the dataset are implicitly labelled. In this research, supervised machine learning technique is used along with ensemble technique.

3.2 Deep Learning

Deep learning is a subset of machine learning which uses neural networks for learning. The neural networks are composed of multiple hidden layers that can capture complex data patterns [13]. The image dataset for analysis has subtle patterns which can be captured by using pre-trained deep learning models specifically CNN.

3.3 Transfer Learning

Transfer learning is harnessing the power of an established model to conquer a fresh challenge. This method entails initially training a model for one task and then using part or all of that model as a foundation for a related task [14]. For instance, if a basic classifier is trained to identify the presence of a backpack in images, the insights gained from that training can be applied to identifying other objects, such as sunglasses. Here are some example models that have been chosen as the base learners for experimental purposes.

3.3.1 ResNet50

ResNet-50 [15] is a highly advanced convolutional neural network which has achieved the best performance in a wide range of computer vision applications, specifically in image classification, object detection, and image segmentation. The architecture has 50 layers, comprising of convolutional layers, pooling layers, and fully connected layers [15]. One of the key ingenuities of ResNet is the integration of residual connections, allowing the network to comprehend residual functions in respect to the inputs of each layer. This makes it easier to train networks with a lot more layers without having to worry about the vanishing gradient. The residual connections enable the network to bypass some layers, allowing information to flow through the network more easily during training.

3.3.2 DenseNet201

DenseNet-201 [16] is a convolutional neural network architecture characterized by dense connections between layers, introduced by the research team at Facebook AI Research. Unlike traditional convolutional neural networks where each layer is connected only to subsequent layers, DenseNet-201 establishes direct connections between all layers, enabling the reuse of features and enhancing the flow of gradient across the network. This architectural design facilitates deeper networks while mitigating vanishing gradient issues, leading to improved learning and feature extraction. Its dense connectivity and efficient parameter sharing make DenseNet-201 a compelling choice for tasks such as object detection, semantic segmentation, and transfer learning in various computer vision applications.

3.3.3 InceptionV3

The InceptionV3 [17] architecture is a highly regarded and implemented CNN, primarily designed for image recognition tasks. It is built upon on its predecessor, InceptionV1, by introducing several enhancements aimed at improving efficiency and performance. Notably, InceptionV3 incorporates factorized convolutions, which reduce computational complexity while preserving representational power, enabling faster training and inference. Additionally, it employs batch normalization and auxiliary classifiers during training to improve convergence and regularization.

3.4 Ensemble Learning

The Ensemble learning is a machine learning paradigm where multiple models, often referred to as "learners" or "base models" are combined to solve a particular problem [18]. The primary goal of ensemble learning is to improve the overall performance, accuracy, and robustness of a predictive model by aggregating the predictions of several models. This approach leverages the idea that while individual models may have different strengths and weaknesses, combining their predictions can result in a more accurate and reliable performance. Weighted Average Ensembles aggregate predictions from numerous models by applying varying weights to each model's forecast [19]. The weights represent the proportional importance or trustworthiness of each model.

3.5 Dataset

For The “Real and Fake Face Detection” dataset provided by Yonsei University available at Kaggle [20] is chosen for this experiment. This dataset comprises a collection of images containing both real and fake human faces. The dataset includes high-quality facial images that were generated by experts. The photos are a collage of various faces split by eyes, nose, mouth, or the entire face. The dataset comprises a total of 2041 images, of which 960 are altered faces and 1081 are genuine faces. A few sample human face images from the dataset [20] are depicted in Fig. 1 which are used in this research work.

Fig. 1: Sample Images From Dataset [20].

4. METHODOLOGY

The following Figure 2 shows the deepfake facial image analysis steps used in this research for real and deepfake human face analysis and detection of the image dataset given in [20].

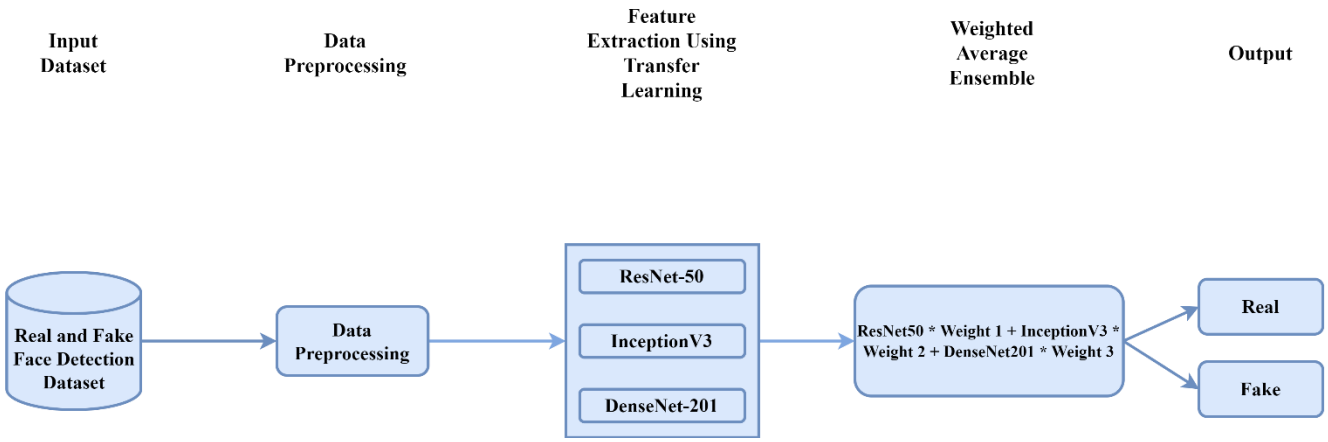

Fig. 2: Deepfake Face Detection Process Using Transfer Learning.

4.1 Dataset Preprocessing

The images in the Real and Fake Face Detection [20] dataset are of 600 x 600 dimension. At first, we resized each image in the dataset to 224 x 224 dimension, and normalization is performed on the images. Then the dataset is splitted into training set, validation set and testing set using a 70:15:15 ratio, resulting 1429 images used for training, 306 images used for validation and 306 images for testing. A batch size of 32 images is also created.

4.2 Model Training

In this research, we have used three pre-trained models: ResNet50, DenseNet201 and InceptionV3 which are originally trained on ImageNet dataset [21]. The output of each base model is enhanced using GlobalAveragePooling2D layer, followed by the ReLU activation function and a dense layer. A dropout layer is employed to avoid overfitting with a rate of 0.4. At the final dense layer of each model, the sigmoid activation function was employed and the corresponding units are expected by our model. In each model, an optimizer is employed as an optimizing technique. Each base model was trained for 20 epochs on the training dataset.

4.3 Weighted Average Ensemble of Three Models

A simple weighted average ensemble [22] approach is selected for experiment purpose. After training, looking at the testing accuracy of the base transfer learning models, we have assigned weight to each model. The testing accuracy of InceptionV3 and DenseNet201 were very close, so they are fixed to 0.5 and 0.4 respectively. The testing accuracy of ResNet50 was much lower than the two other models, it was fixed at 0.1. A user defined function WeightedAvgLayer is defined to perform the weighted average task which takes all of the 3 weight values as the argument. Finally, the ensemble model is trained for 10 epochs on the training data set.

5. EXPERIMENTAL RESULT

5.1.1 Result Analysis

The following Table 1 shows the efficiency of the 3 pre-trained models.

Table 1. Accuracy and Loss comparison of different model

Model	Val_acc	Val_loss	Test_acc	Test_loss
ResNet50	0.5359	0.6913	0.5359	0.6897
DenseNet201	0.6503	0.6710	0.6209	0.7173
InceptionV3	0.6013	0.6657	0.6340	0.6744
Weighted Average Ensemble	0.5948	0.7156	**0.6471**	0.7312

We can see a clear increase in the testing accuracy after ensembling three models.

The following Table 2 shows performance metrics of different models.

Table 2. Performance Metrics

Model	Accuracy	Precision	Recall	F1-score
ResNet50	0.54	0.56	0.51	0.39
DenseNet201	0.62	0.62	0.62	0.62
InceptionV3	0.63	0.63	0.63	0.63
Weighted Average Ensemble	0.65	0.65	0.64	0.64

From the table it is clearly visible that among all the models, ensemble has the slightly improved value in all criteria.

From the confusion matrix in the following Figure 3(a) we can see that among 306 images in the training set 5 real images are predicted as fake and 137 fake images are predicted as real. The confusion matrix for densenet201 in Figure 3(b) depicts that 64 real images are predicted as fake and 52 fake images are predicted as real. In Figure 3(c) we can see 42 real images are predicted as fake and 70 fake images are predicted as real. The confusion matrix in Figure 3(d) is the ensemble one, where we can see 41 real images are predicted as fake and 67 fake images are predicted as real

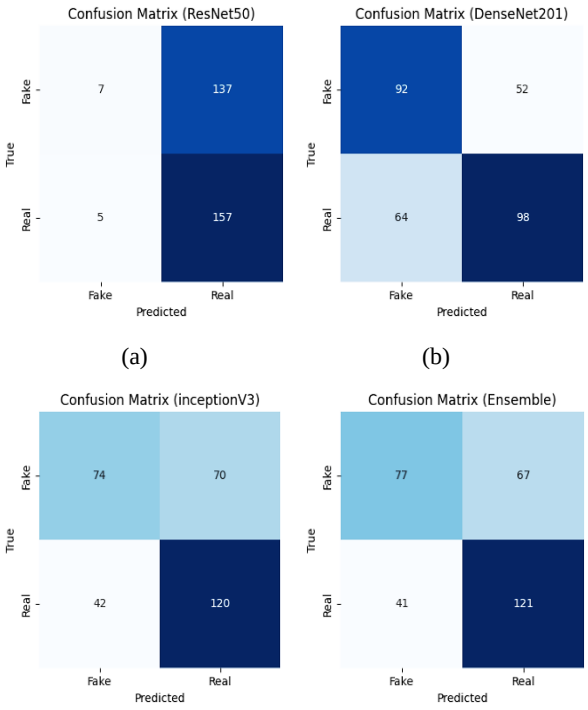

Fig. 3: Confusion matrix of different models.

5.1.2 Performance Comparison

Performance comparison with other state of the art methods is given below in Table 3.

Table 3. Accuracy Comparison

Reference	Year	Dataset	Techniques Used	Accuracy (%)
Suratkar et al [22]	2020	FaceForensics++ + Google AI	VGG16, DenseNet121, XceptionNet, InceptionV3	95
Suratkar et al [23]	2020	FaceForensics++ + Google Deep Fake Dataset + Deep Fake Dataset from facebook	InceptionV3, ResNet50, ResNet18, NasNetMobile, MobileNet, Xception	91.5
Patel et al [24]	2020	Deepfake Detection Challenge (DFDC)	VGG16, DenseNet121, MobileNet, InceptionV3, ResNet50	90.2
Raza et al [2]	2022	Real and Fake Face Detection	Novel DFP	94
Mahmud et al [25]	2023	FaceForensics++ and CelebDf-V2	XceptionNet, InceptionResNetV2, EfficientNetV2S, EfficientNetV2M	98 and 94
Coccomini et al [26]	2022	DFDC + FaceForensics++	EfficientNetB0, Vision Transformer	80
Gong et al [27]	2021	FaceForensics++, TIMITand Kaggle competitions	DeepfakeNet	96.69
Atwan et al [28]	2024	deepfake and real images	Xception, DenseNet121, MobileNet, InceptionV3, ResNet50	86.58
This Work	2024	Real and Fake Face Detection	Weighted Average Ensemble	64.71

6. CONCLUSION

In this work, we employed transfer learning and weighted average ensemble technique to detect fake human faces. We have used three different pre-trained models namely ResNet50, Dense201 and InceptionV3. The reason of using pre-trained models is that transfer leaning minimizes training time as the model is not trained from scratch. Also, we have combined the prediction from three models using weighted average ensemble. It significantly improved the accuracy of the ensemble model than the individual models. This study significantly addresses the extent of success that can be achieved in detecting manually altered deepfake images using the aforementioned approach. In future, we intend to include latest models in the ensemble model and evaluate their performance in the automatic detection of fake human faces. Also the research can be continued including diverse datasets to improve the accuracy.

7. ACKNOWLEDGEMENT

The research work presented in this paper is an outcome of the ongoing M.Sc. thesis work of the first author which is funded by the NST fellowship under the Ministry of Science and Technology, Dhaka, Bangladesh in the fiscal year 2023-2024. We also thank the experts and personnel in relation to this research work.

8. REFERENCES

[1] Khichi, M., and Yadav, R. K. (2021, December). Analyzing the methods for detecting deepfakes. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 340-345). IEEE.

[2] A.Raza, K.Munir, and M.Almutairi, “A novel deep learning approach for deepfake image detection,” Applied Sciences, vol. 12, no. 19, p. 9820, 2022.

[3] Y. Abdalla, M. Iqbal, and M. Shehata, “Image forgery detection based on deep transfer learning,” European

Journal of Electrical Engineering and Computer Science, vol. 3, no. 5, 2019.

[4] N. Kumar, P. Pranav, V. Nirney, and V. Geetha, "Deepfake image detection using cnns and transfer learning," in 2021 International Conference on Computing, Communication and Green Engineering (CCGE). IEEE, 2021, pp. 1–6.

[5] X. Chang, J. Wu, T. Yang, and G. Feng, "Deepfake face image detection based on improved vgg convolutional neural network," in 2020 39th chinese control conference (CCC). IEEE, 2020, pp. 7252–7256.

[6] Patrick Schneider, and Fatos Xhafa, in Anomaly Detection and Complex Event Processing over IoT Data Streams, 2022

[7] Qureshi, A.S., and Roos, T. (2021). Transfer Learning with Ensembles of Deep Neural Networks for Skin Cancer Detection in Imbalanced Data Sets. Neural Processing Letters, 55, 4461 - 4479.

[8] Vallabhajosyula, S., Sistla, V., Kolli, and V.K. (2021). Transfer learning-based deep ensemble neural network for plant leaf disease detection. Journal of Plant Diseases and Protection, 129,545 - 558.

[9] Sharma, J., Sharma, S., Kumar, V., Hussein, H.S., and Alshazly, H.A. (2022). Deepfakes Classification of Faces Using Convolutional Neural Networks. Traitement du Signal.

[10] Shad, H.S., Rizvee, M.M., Roza, N.T., Hoq, S.M., Khan, M.M., Singh, A., Zaguia, A., and Bourouis, S. (2021). Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network. Computational Intelligence and Neuroscience, 2021.

[11] Rana, M.S., Nobi, M.N., Murali, B., Sung, and A.H. (2022). Deepfake Detection: A Systematic Literature Review. IEEE Access, 10, 25494-25513.

[12] Deep Learning Approaches for Early Diagnosis of Neurodegenerative Diseases.DOI:10.4018/979-8-3693-1281-0.ch011

[13] https://www.ibm.com/topics/deep-learning

[14] Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[15] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[16] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[17] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[18] https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

[19] Mohammed, A., and Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 35(2), 757-774.

[20] Available: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection

[21] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[22] Suratkar, S., Kazi, F., Sakhalkar, M., Abhyankar, N., and Kshirsagar, M. (2020, December). Exposing deepfakes using convolutional neural networks and transfer learning approaches. In 2020 IEEE 17th India council international conference (INDICON) (pp. 1-8). IEEE.

[23] Suratkar, S., Johnson, E., Variyambat, K., Panchal, M., and Kazi, F. (2020, July). Employing transfer-learning based CNN architectures to enhance the generalizability of deepfake detection. In 2020 11th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-9). IEEE.

[24] Patel, M., Gupta, A., Tanwar, S., and Obaidat, M. S. (2020, October). Trans-DF: a transfer learning-based end-to-end deepfake detector. In 2020 IEEE 5th international conference on computing communication and automation (ICCCA) (pp. 796-801). IEEE.

[25] Mahmud, F., Abdullah, Y., Islam, M., and Aziz, T. (2023, December). Unmasking Deepfake Faces from Videos Using An Explainable Cost-Sensitive Deep Learning Approach. In 2023 26th International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[26] Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In International conference on image analysis and processing (pp. 219-229). Cham: Springer International Publishing.

[27] Gong, D., Kumar, Y. J., Goh, O. S., Ye, Z., and Chi, W. (2021). DeepfakeNet, an efficient deepfake detection method. International Journal of Advanced Computer Science and Applications, 12(6), 201-207.

[28] Atwan, J., Wedyan, M., Albashish, D., Aljaafrah, E., Alturki, R., and Alshawi, B. (2024). Using Deep Learning to Recognize Fake Faces. International Journal of Advanced Computer Science & Applications, 15(1).