International Conference on Wireless Technologies, Networks and Science-2025 (Hybrid Mode): (May 10-11, 2025)

ISBN Number: 978-93-340-1708-3

Binary Classification of Facial Features for Robust Deepfake Identification

Gaurav Aggarwal[1], Gaurav Sharma[2], Pooja Joshi[3], Ashutosh Bhatt[4], Parul Saini[5]

[1]Swami Rama Himalayan University, Dehradun, India,gauravaggarwal@srhu.edu.in
[2]Swami Rama Himalayan University, Dehradun, India, gauravsharma@srhu.edu.in
[3] Swami Rama Himalayan University, Dehradun, India, poojajoshi.baloni@gmail.com
[4] Swami Rama Himalayan University, Dehradun, India, ashutoshbhatt15@gmail.com
[5]DIT University, Dehradun, India, parul.saini@dituniversity.edu.in

ABSTRACT:; Deepfake technology has emerged as a significant challenge in the realm of digital media due to its potential misuse in misinformation, privacy violations, and security breaches. This study focuses on the binary classification of facial features as a means of enhancing Deepfake identification. By using the natural differences between real and fake facial features, we proposed a strong framework that uses deep learning models to look at small differences in texture, lighting, and facial movement. Our approach involves pre-processing facial data to isolate key regions prone to manipulation, such as the eyes, mouth, and skin texture. We employ multiple convolutional neural networks (CNNs) and dense layers to extract high-dimensional features, subsequently classifying them into two categories: real or fake. To improve robustness, the model incorporates data augmentation and ensemble techniques to handle variations in quality and environmental factors. We validate the framework on the Kaggle dataset, achieving state-of-the-art accuracy while maintaining computational efficiency. Our results demonstrate the potential of the binary classification of facial features as a reliable and scalable solution for deepfake detection, paving the way for practical applications in media forensics and digital security.

Keyword: Deepfake image, CNN, Deepfake detection

1. INTRODUCTION

In the digital age, where pictures and videos are the main form of communication, the authenticity of visual content has emerged as a crucial component of credibility and trust. Particularly important is facial imagery, which has an effect on a wide range of industries, including law enforcement, social media, entertainment, and journalism. However, the integrity of visual content is now being threatened by the growth of synthetic media, particularly deepfake technology. Deepfakes have raised worries about false information, fraud, and harmful activity since they use sophisticated artificial intelligence to create incredibly lifelike fake face content. Therefore, maintaining the integrity of facial features in digital media is crucial to preserving social stability and public confidence [1].

The challenges posed by image manipulation, especially through deepfake technology, are vast and multifaceted. Unlike traditional methods of photo and video editing, deepfakes utilize neural networks to fabricate highly convincing synthetic content, making them exceptionally difficult to detect with the naked eye. This has led to widespread misuse, including the creation of fake news, identity theft, and other forms of cybercrime. Moreover, as manipulation techniques grow increasingly sophisticated, the gap between genuine and manipulated content continues to narrow, further complicating the detection process [1].

There has never been a greater need for efficient and scalable detection techniques given the seriousness of these issues. Deepfakes have been difficult to spot using conventional detection methods like manual inspection or metadata analysis, especially as modification tools have advanced. AI and machine learning-based automated methods have become a viable remedy. These techniques seek to offer reliable mechanisms for differentiating authentic content from fake media by spotting minute irregularities in facial features, motion, or texture [2].

The binary categorization of face features as a unique method for reliable deepfake identification is the main emphasis of this work. This study focuses on particular facial irregularities and discrepancies that are created throughout the deepfake generation process, in contrast to many other approaches that use broad-spectrum analysis. The work aims to increase the precision and dependability of deepfake detection systems by focusing on fine-grained patterns in facial geometry, texture, and emotion [3].

Although the industry has made significant strides, there are still a number of issues with the models that are now in use. A lot of the detection methods that are used today have problems like not being able to work with all datasets, being slow to process, and being easy for attackers to break into. Furthermore, most models fail to consider small face traits, which frequently serve as warning indicators of modification. These gaps highlight the necessity for specialized, lightweight, and scalable systems that can handle the particular difficulties presented by deepfakes [3].

This study offers two contributions. In order to identify deepfakes, it first presents a binary classification system that focuses on spotting facial discrepancies. This method uses state-of-the-art deep learning architecture to accurately assess facial traits. Finally, the paper shows how the model works on different datasets to show that the suggested approach is useful for delivering reliable and scalable deepfake detection. This work helps to develop useful tools for countering the growing threat of deepfake technology by filling important gaps in the field.This study aims to make the field of visual content authentication better by looking into the problems of deepfake identification. This will help make sure that digital media is reliable and honest in a time when fake content is becoming more common.

2. LITERATURE REVIEW

The increasing prevalence of deepfake technology has spurred significant research into the development of effective

ISBN Number: 978-93-340-1708-3

detection methods, with a focus on convolutional neural networks (CNNs) and advanced machine learning techniques. A study on deepfake binary facial image classification shows that custom CNN models, along with ResNet-50 and EfficientNet B7 transfer learning frameworks, can accurately spot content that has been changed. This research emphasizes the critical need for robust detection methods to mitigate the growing digital security risks posed by deepfakes [4].

Ajoy et al. [5] propose a CNN-based model to detect deepfakes by identifying distinct features like pixel distortions and facial inconsistencies. The model aids social media platforms in mitigating the spread of harmful fake videos and preventing misinformation and unrest. Another study redefines deepfake detection as a fine-grained classification problem, moving beyond the traditional binary classification paradigm. The proposed method improves robustness across datasets and manipulation types by learning subtle, generalizable features and blocking background noise. This fixes some of the main problems with current methods [6].

Jolly et al. [7] describe an automated way to find facial expressions in videos using Deepfake, Face2Face, faceSwap, and neural texture. They do this by using a layered approach that gets around problems with data loss and compression. Building on CNN-based models, another investigation leverages generative adversarial networks (GANs) and data augmentation for dataset creation. Using pretrained models such as VGG16 and ResNet-50, the study achieves notable accuracy improvements, particularly through ensemble methods, which combine these models to achieve accuracies as high as 98.79% on benchmark datasets [8].

Altaei et al. [9] use CNN to detect fake images. It involves preprocessing, Gamma correction, and Canny filter extraction. We use two detection methods: CNN with Principal Component Analysis (PCA) and CNN without PCA. Results show acceptable accuracy, but CNN only provides the highest accuracy. Finally, an innovative approach blends scalable CNNs, such as EfficientNet, with hierarchical vision transformers (ViT), specifically the shifted window transformer. This hybrid architecture achieves a remarkable accuracy of 98.04%, demonstrating the potential of combining CNNs with transformers for enhanced deepfake detection [5].

3. PROPOSED DEEPFAKE DETECTION

This section explains the proposed methodology for deepfake image identification. An essential resource for deepfake detection, especially for binary classification using Convolutional Neural Networks (CNNs), is a dataset of approx. 190,000 real and fake facial photos. Table 1 contains all the information about the dataset used. The collection of real and altered photos in this dataset is well-balanced and varied, capturing differences in lighting, facial expressions, and alteration methods.Several qualities are necessary for training CNNs to properly generalize across real-world contexts.

Since CNNs are so efficient at learning hierarchical features, they can identify minute irregularities brought forth by deepfake creation, like variations in texture, lighting, and face alignment. Researchers can create strong solutions to combat deepfake issues by using this dataset to thoroughly train and assess detection methods. These developments counter the transmission of false information and fortify digital media authentication systems. Figure 1 illustrates the implementation of the dataset.

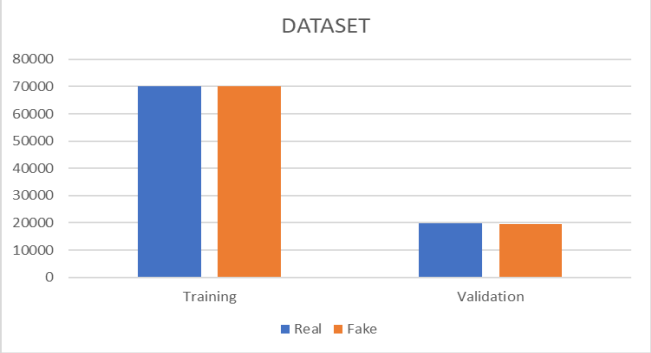

Figure 1: Implementation details of the Dataset

Table 1: Used Dataset in Proposed model

Dataset	Images	Resolution
190K Fake and Real Faces	190000	256*256

The proposed model architecture is shown in Figure 2. The proposed model leverages a CNN architecture specifically designed for deep fake image detection. This model is structured to extract and analyze spatial features in images to

ISBN Number: 978-93-340-1708-3

classify them as real or fake effectively.The input to the model is an RGB image with a fixed size of 256x256 pixels. The initial layers of the model consist of a series of Conv2D layers, each equipped with 128 filters and 3x3 kernels. These layers extract low-level features such as edges, textures, and simple patterns. Each convolutional operation is followed by a batch normalization layer, which normalizes the activations and accelerates the training process.

We incorporate Max Pooling layers to further reduce the computational complexity, downsampling the feature maps by selecting the most significant values from local regions.The model includes four convolutional blocks, each progressively reducing the spatial dimensions of the feature maps. Starting from an input size of 256x256, the feature maps are reduced to 14x114×14. This dimensionality reduction allows the model to focus on high-level spatial features while maintaining computational efficiency. The output of the convolutional blocks is then passed through a Flatten layer, which converts the 3D feature map into a 1D vector of size 25088.

This transformation prepares the data for the fully connected layers, enabling classification.The flattened features are processed through multiple dense layers, each containing 128 neurons. To prevent overfitting and enhance generalization, Dropout layers are added between the dense layers. These layers randomly deactivate a fraction of the neurons during training, ensuring the model does not overly rely on specific features. The final dense layer contains a single neuron, which outputs a binary value indicating whether the input image is real or fake.The model comprises approximately 3.8 million parameters, of which 3.7 million are trainable.

This balance between complexity and efficiency allows the model to achieve high accuracy in detecting deepfake images while remaining computationally feasible. The combination of convolutional, pooling, and dense layers ensures that the model captures both local and global features, making it robust against variations in deep fake techniques.

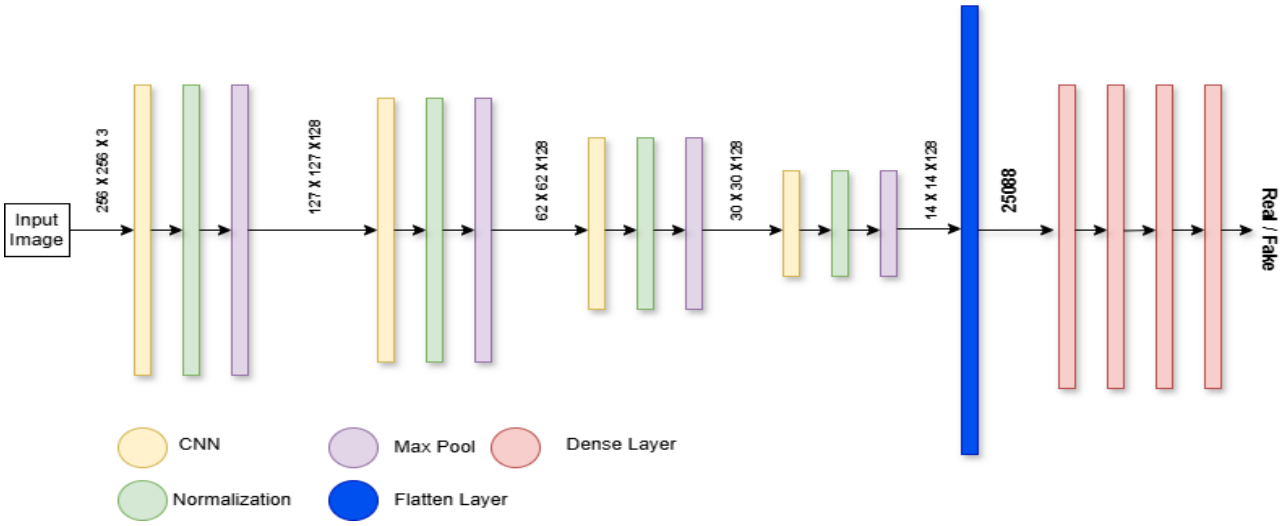

Figure 2: Architecture of Proposed Model for Deep Fake Image detection

ISBN Number: 978-93-340-1708-3

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 254, 254, 128)	3,584
batch_normalization (BatchNormalization)	(None, 254, 254, 128)	512
max_pooling2d (MaxPooling2D)	(None, 127, 127, 128)	0
conv2d_1 (Conv2D)	(None, 125, 125, 128)	147,584
batch_normalization_1 (BatchNormalization)	(None, 125, 125, 128)	512
max_pooling2d_1 (MaxPooling2D)	(None, 62, 62, 128)	0
conv2d_2 (Conv2D)	(None, 60, 60, 128)	147,584
batch_normalization_2 (BatchNormalization)	(None, 60, 60, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 30, 30, 128)	0
conv2d_3 (Conv2D)	(None, 28, 28, 128)	147,584
batch_normalization_3 (BatchNormalization)	(None, 28, 28, 128)	512
max_pooling2d_3 (MaxPooling2D)	(None, 14, 14, 128)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 128)	3,211,392
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 128)	16,512
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 128)	16,512
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 128)	16,512
dropout_3 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 1)	129

Total params: 3,709,441 (14.15 MB)
Trainable params: 3,708,417 (14.15 MB)
Non-trainable params: 1,024 (4.00 KB)

Figure 3: Summary of the Architecture of Proposed Model for Deep Fake Image detection

4. EXPERIMENTS AND RESULTS

The proposed model performs exceptionally well on a dataset of 190,000 actual and fraudulent face photos, as Tables 2 and 3 illustrate. With a training accuracy of 99.23%, the model demonstrated a great capacity for learning from the training data. Its 94.5% F1 score ensures dependable classification performance by reflecting a balanced trade-off between precision and recall.

Table 2: Performance comparison of proposed model and other approaches used Dataset

Model name	Test Accuracy (%)
ResNet50 [11]	53.43
FaceNet [12]	94.51
Conv2D CNN [8]	95.85
Proposed Model (CNN)	99.23

Table 3: Performance evaluation of the proposed model

Model	Dataset	Train Accuracy	F1 Score	Precision	Recall
Proposed Model	190k Real and Fake Faces	99.23	94.5	95	94

ISBN Number: 978-93-340-1708-3

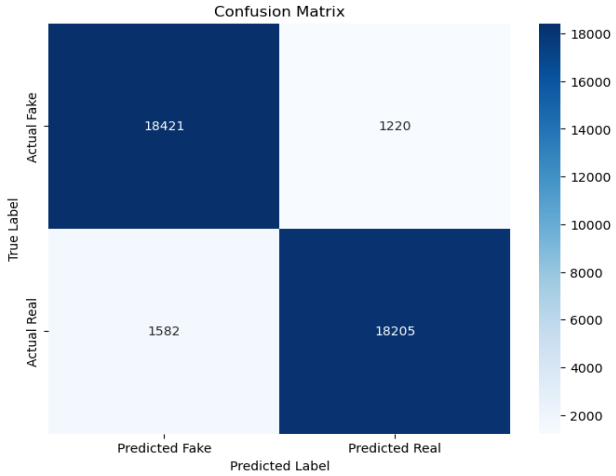

Figure 4: Confusion Matrix

With a precision of 95%—a metric that quantifies the ratio of true positives to anticipated positives—the model is clearly effective at detecting phony faces with little false positives. Its reliability in differentiating between real and phony faces is further demonstrated by its 94% recall score, which shows how well it detects actual affirmative situations. These outcomes highlight how well the model handles this difficult categorization task and how robust it is.

From Table 2 and Table 3, following observations have been made:

- The proposed model for deep fake image detection demonstrates exceptional performance when compared to existing models, achieving a test accuracy of 99.23%.
- A high train accuracy (99.23%) confirms the effectiveness of the architecture in learning meaningful patterns.
- The F1 score (94.5) ensures balanced performance between precision and recall.
- The high precision (95%) reduces false positives, and the high recall (94%) ensures effective identification of fake images.

The confusion matrix for the binary-class classification, which includes Fake and Real, is shown in Figure 4. A confusion matrix of dimension 2 × 2 (for binary classification) linked to a classifier displays the expected and actual classification, where 2 is the number of unique classes. The true positives and true negatives in the diagonal elements for the suggested model are represented by the confusion matrix on the 190k Real and Fake Faces dataset.

5. CONCLUSION

The proposed model provides a comprehensive approach to deep fake detection by leveraging the power of CNNs to identify intricate patterns and anomalies.The proposed CNN-based model achieves state-of-the-art performance in deep fake image detection, with an accuracy of 99.23%, demonstrating significant improvements over ResNet50, FaceNet, and other Conv2D CNN approaches. Its superior ability to extract and analyze features tailored to detecting deep fake artifacts highlights the importance of designing specialized architectures for this domain. The inclusion of batch normalization, dropout, and multiple dense layers ensures that the model is both accurate and resilient to overfitting, making it a strong candidate for real-world applications in detecting manipulated media.

REFERENCES

[1] 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. https://doi.org/10.1016/j.inffus.2020.06.014.

[2] Khalil, H.A., Maged, S.A. (2021). Deepfakes creationand detection using deep learning. In 2021 InternationalMobile, Intelligent, and Ubiquitous ComputingConference (MIUCC), pp. 1-4.https://doi.org/10.1109/MIUCC52538.2021.9447642Detecting Facial Image Manipulations with Multi-Layer CNN Models.

[3] Nguyen, T., Nguyen, Q.V.H., Nguyen, C.M., Nguyen, D.,Nguyen, D.T., Nahavandi, S. (2019). Deep learning fordeepfakes creation and detection: A survey. arXiv preprintarXiv:1909.11573.https://doi.org/10.48550/arXiv.1909.11573.

[4] M. Kalemullah, P. Prakash and V. Sakthivel, "Deepfake Classification For Human Faces using Custom CNN," 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2024, pp. 744-750, doi: 10.1109/ICCPCT61902.2024.10672973.

[5] A. Ajoy, C. U. Mahindrakar, D. Gowrish and V. A, "DeepFake Detection using a frame based approach involving CNN," 2021 Third International Conference on

Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1329-1333, doi: 10.1109/ICIRCA51532.2021.9544734.

[6] A. V. Nadimpalli and A. Rattani, "Facial Forgery-Based Deepfake Detection Using Fine-Grained Features," 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, 2023, pp. 2174-2181, doi: 10.1109/ICMLA58977.2023.00328.

[7] V. Jolly, M. Telrandhe, A. Kasat, A. Shitole and K. Gawande, "CNN based Deep Learning model for Deepfake Detection," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-5, doi: 10.1109/ASIANCON55314.2022.9908862.

[8] J. Sharma, S. Sharma, V. Kumar, Hany S. Hussein, and H. Alshazly "Deepfakes Classification of Faces Using Convolutional Neural Networks". Traitement du Signal 39, 3 (2022), 1027–1037. DOI:http://dx.doi.org/10.18280/ts.390330.

[9] Altaei, Mohammed Sahib Mahdi. "Detection of Deep Fake in Face Images Using Deep Learning." Wasit Journal of Computer and Mathematics Science 1, no. 4 (2022): 60-71.

[10] Kerenalli, S., Yendapalli, V., Mylarareddy, C. (2023). Fake Face Image Classification by Blending the Scalable Convolution Network and Hierarchical Vision Transformer. In: Reddy, K.A., Devi, B.R., George, B., Raju, K.S., Sellathurai, M. (eds) Proceedings of Fourth International Conference on Computer and Communication Technologies. Lecture Notes in Networks and Systems, vol 606. Springer, Singapore. https://doi.org/10.1007/978-981-19-8563-8_12

[11] [11]Hettiarachchi, S. (2021). Analysis of different face detection and recognition models for Android. Digitala Vetenskapliga Arkive

[12] Wen L., Xu, D. (2019). Face image manipulation detection. In IOP Conference Series: Materials Science and Engineering, 533(1): 012054. https://doi.org/10.1088/1757-899X/533/1/012

.