IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Efficient And Explainable Deepfake Detection System Using Deep Learning

Prabha M, Sahana K, Swedha S, and Shobhanjaly P Nair

Department of Computer Science and Technology, Sri Venkateswara College of Engineering, Chennai, India.

Doi: https://doi.org/10.56975/ijcrt.v14i7.309009

Abstract Generative AI's development to the point where synthetic media can be so close to actual media that it raises issues; e.g. when people use manipulated video clips of public figures to spread false information (deepfakes), they create confusion regarding truthfulness and decrease reliability for all forms of electronic information. This research explains the development

of a method to deal with this issue both effectively and practically, by constructing an automated system that not only detects deepfakes with high accuracy, but also provides an explainable reason(s) as to why it categorizes specific items as deepfakes. This method also includes the use of machine learning-based algorithmic feature extraction for building the overall framework and machine learning algorithms for determining the explainable results. The overall framework and method have a very defined process from data collection, isolation of facial areas, preprocessing images and finding features of the images via convolutional neural networks and then classifying the images. During this process, the model will focus on finding those minute discrepancies (slight distortions in texture and unnatural blending) that are the signatures of deepfake models even if they are convincing to people. The features that are extracted during the feature extraction step are then compressed to form a smaller, more manageable representation and sent to the classifier to determine if the image is real or manipulated; however, just using accuracy was not enough for us and we also added an explainability layer by using Grad-CAM so that we could view which areas the model is focusing on. Sometimes it is the eyes, sometimes it is the jawline and, interestingly, it is not always what a human would look at first. Ultimately, the goal of this entire project is to not only detect deepfakes, but to understand why they are detected as deepfakes so that we can develop systems that will be able to eliminate the possibility of deepfakes in the future.

Keywords: Deepfake Detection, Explainable AI, Grad-CAM, SHAP, LIME, CNN, Transformer, InceptionV3, Xception, CSWin Transformer.

1 Introduction

The progress in artificial intelligence, especially in deep learning, has significantly simplified the process of creating synthetic media [6]. For instance, a few years ago, it took a lot of work to create a fake video of a person. Today, however, with tools like GANs, it is possible to create a fake video with tools that are becoming increasingly accessible [4].

Perhaps the most interesting example of synthetic media is deepfakes. Using deepfakes, it is possible to replace a person's face, voice, or expressions with such convincing results that even a careful viewer can be misled [3]. For instance, in a movie
or a game, it can be useful. However, outside a controlled environment, things get complicated. Fake political speeches, fake news clips, and identity theft are no longer hypothetical scenarios [6].

Things are made worse by how fast synthetic media can spread. For instance, a fake video uploaded to social media does not remain there. Instead, it spreads. By the time a person is suspicious of it, it is already too late. Traditional approaches, such as checking for compression artifacts or pixel-level inconsistencies, were effective to a certain extent [3]. However, with modern deepfake tools, it seems they have become much better at concealing those [4]. Hence, it is only logical to move on to deep learning approaches, as they can detect inconsistencies that are not immediately apparent to the naked eye [1].

However, there is a catch. Most of these approaches are black boxes. They will provide an answer but will never provide a reason for it [4]. Now, it is true that sometimes a black box approach is necessary. But in cases where it is used in journalism or as part of a court trial, it is certainly not advisable.So, instead of focusing purely on accuracy, there is a move towards a different approach. The approach is a combination
of convolutional neural networks and transformer-based neural networks [1]. The idea is to ensure a balance between performance and explainability. The former is achieved through the use of explainable AI [4]. The balance between performance and explainability is certainly not easy to achieve.

2 Literature Review

Magesh et al. [1], "Building an Efficient DeepFake Detection System Using CNNs and Transformers" (2025). It is the most relevant piece of research in terms of the present study. Magesh et al. argue that the combination of preprocessing techniques relying on CNNs with the CSWin Transformer helps learn both local and global deepfake characteristics at once without distinguishing them as different features [1]. The efficiency of their method proves impressive: the accuracy of the model reaches 98.7%, while the F1-score

amounts to 98.72%. This paper has demonstrated that with proper configuration, transformers can easily outperform CNNs [1].

The relevance of the chosen paper lies in the fact that the proposed hybrid architecture based on the integration of InceptionV3, Xception, and the CSWin Transformer is directly inspired by this work [1]. It is clear that the necessity to take into account local and global deepfake artifacts proved evident. Without this knowledge, the decision to rely on the specified design would have seemed unfounded and arbitrary.

Khan, Shahzad & Syed (2025), "A Survey on Multimedia-Enabled Deepfake Detection" [6]. While some surveys may end up being an amalgamation of summaries, this particular paper has been found quite informative. Khan et al. discuss the detection of deepfakes with regard to images, videos, audio, and multimodal media types, addressing their shortcomings with regards to generalization and interpretability [6]. It turns out that the failure of detection systems with respect to generalizing performance is something most studies conveniently skip, focusing solely on benchmark tasks. Additionally, the authors mention explainable AI, federated learning, and blockchain technology as potentially interesting research directions [6].

The most valuable part of this paper was its perspective on the topic. The fact that current research requires the incorporation of explainability and does not limit itself to binary predictions directly ties into the gap identified in this survey paper [6]. Additionally, the limitation with regards to generalization was also discussed by the authors. All in all, the paper offered insight into the insufficiency of relying solely on accuracy for evaluation.

Singh, Charanarur, and Chaudhary (2025), "Advances in the Detection of Deepfakes: AI Algorithms and Future Perspectives" [7]. In contrast to Khan et al., this review article takes a slightly different direction, analyzing algorithm-based comparisons of convolutional networks, long-short-term memory networks, and other deep learning models used for image, video, and audio-based deepfake detection [7]. Their key message is that despite the considerable success achieved in improving the detection process through the use of machine learning, none of the algorithms can detect deepfakes in all media types equally effectively [7].

The usefulness of this study for the current research project is related to its comparative analysis. When selecting the type

of architecture, it is helpful to have an idea about the proven effectiveness of certain families of models on standard benchmark tests [7]. The choice of using convolutional neural networks, such as Xception and InceptionV3, was partly based on this information. In general, the paper made a strong case for remaining within the framework of deep learning rather than feature engineering.

Al Kishri et al. (2025), "A Comparative Study of Deepfake Facial Manipulation Techniques Using GANs" [5]. This paper steps back and looks at the issue from the perspective of face generation, discussing how deepfake technologies using GANs

operate and how detection systems work to counter them [5]. Al Kishri et al. discuss various detection methods

that have been used before, including handcrafted, artifact, and learning-based approaches, analyzing GAN architectures and available benchmark datasets for commonalities and, more importantly, shortcomings [5]. Their takeaway is that the current benchmarking approach shows significant bias and poor practical application, with GAN-independent techniques required for future research [5].

It goes without saying that any paper discussing detection must give at least some thought to how deepfake images are created in the first place. This paper provides some much-needed context for the current study and provided the necessary basis for understanding the features of the GAN-generated faces that the detection algorithm needed to detect [5]. The criticism that the benchmarks themselves are biased and unrealistic is mentioned in the discussion section of this work, in part due to the realization brought by this paper.

3 Related Work

Deep learning-based deepfake detection was not the first method, and it was much simpler in the beginning, relying on the detection of obvious differences, such as differences in lighting, textures, or compression [3]. These methods worked quite well in the beginning, but as the quality of fake media improved, these differences began to disappear [3].

This is where CNN-based methods came into the picture, and methods such as Xception and EfficientNet have proven that it is possible to detect deepfakes by learning spatial patterns in images [1], [7]. The CNN does not require any prior knowledge

of what a fake looks like; it learns by itself and often does a much better job [1].

Recently, transformer-based methods have also come into the picture, and these methods are quite different from CNN- based methods, as they do not look at a part of the image individually, as CNNs do, but take a broader view of the entire image and try to detect differences in the behavior of different parts of the face, such as the eyes and the mouth, and methods such as Vision Transformers and CSWin Transformers are quite useful for this purpose [1], [7]. There's also been a growing interest in explainability. SHAP, LIME, and Grad-CAM are tools designed to help answer a simple question: why did the model make that decision? [4] The results are sometimes interesting. Other times, a bit surprising.

Despite all these advances, there are still some challenges. For example, a model built on one data set sometimes doesn't work on another. The computational cost is still high [6]. And interpretability is still a work in progress. So there is certainly room for improvement.

4 Methodology

The methodology is organized into a series of clearly defined steps, with each step fulfilling a particular role. The complete structure of the suggested system is illustrated in Figure 1.

4.1 Data Preprocessing

Here, the data is in the form of images, both real and fake. Faces are detected and aligned using MTCNN [1]. Then, these faces are cropped, resized, and normalized. It sounds like a mundane task, but it is quite important and makes a big difference later on. Some augmentation is done, like flipping and rotation, to prevent overfitting.

4.2 Feature Extraction

Now, the actual work begins. Instead of using a single model, a combination is used. CNNs like InceptionV3 and Xception look at the small, local features, like small irregularities that might occur at the edges [1], [7]. Then, a CSWin Transformer tries to look at the larger picture and identify the inconsistencies in the entire face. Finally, the output is combined and reduced using a technique called Global Average Pooling. This is because the output is not too large, but it is still quite informative.

4.3 Classification

The feature vector is fed into a fully connected layer. The Softmax function is used to classify the image as real or fake. The training process utilizes categorical cross entropy, which enables the model to learn end-to-end.

4.4 Explainability

In this part, the model goes one step further than just predicting the result. The explainability of the result is obtained through SHAP or LIME [4], which highlights the parts of the image used to make the prediction. At times, the result makes sense to human intuition. At other times, it doesn't, which, in a way, speaks volumes.

The result is quite simple: "Real" or "Fake," along with a visual explanation.

5 Results and Experiments

The deep learning components of the suggested system are managed using TensorFlow and Keras, which are created in Python. The datasets used to evaluate the model are divided into training, validation, and testing groups and contain both genuine and false images.

The model gains the ability to identify both little defects and general irregularities in deepfake photos throughout training. Common metrics including accuracy, precision, recall, and F1-score are used to evaluate the model's performance. Figure 2 displays the ROC curve that illustrates the model's performance.

Strong findings from the model demonstrate that the hybrid approach outperforms the use of a single model in identifying manipulation patterns. The accuracy of the system's classification is further demonstrated by the confusion matrix in Figure 3.

Interpretability is crucial in this work in addition to performance. The model provides graphic explanations

that highlight the factors that influenced its decision. Figures 4 and 5 show the explainability results from SHAP and LIME, respectively.

Traditional deep learning models frequently overlook this additional layer of explainability, which fosters trust and provides a clearer picture of the model's decision-making process.

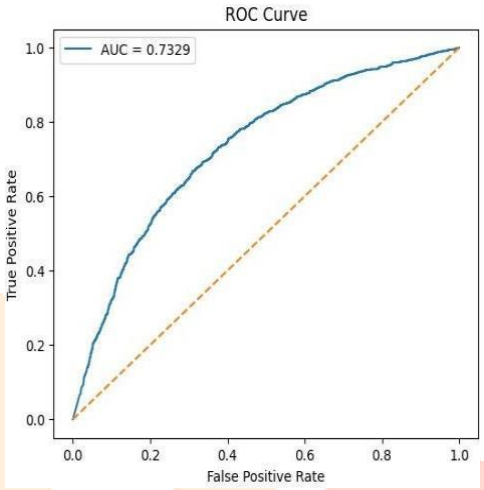

Figure 2: ROC curve graph.

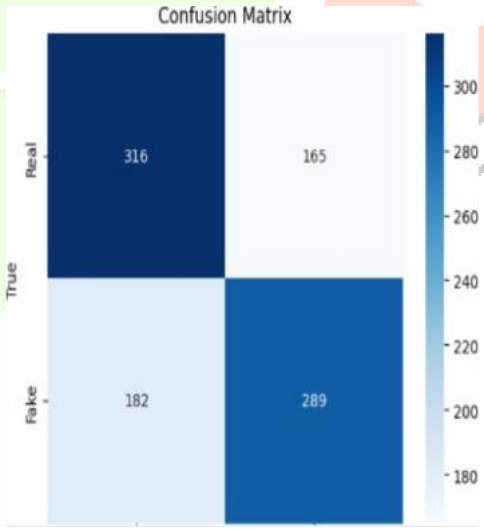

Figure 3: Confusion matrix of our model.

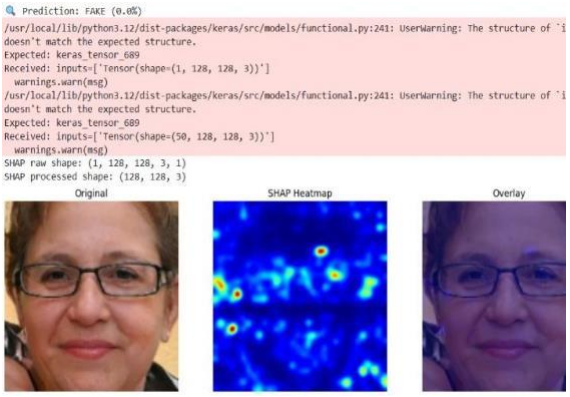

Figure 4: Explainable AI (SHAP) Output.

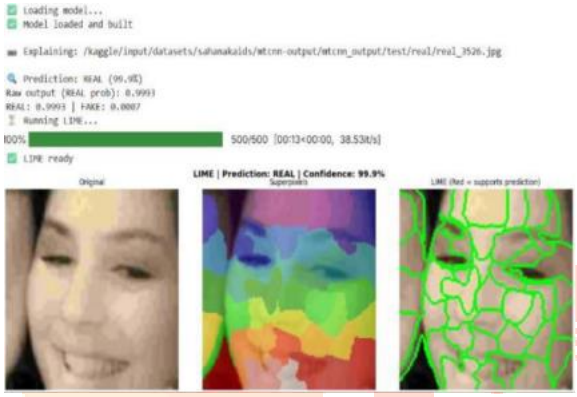

Figure 5: Explainable AI (LIME) Output.

6 Application

The suggested deepfake detection system has various applications in different areas:

Social Media Monitoring– Detection of fake images and videos before they go viral on various social media platforms. Digital Forensics– It can aid the investigation team in verifying the authenticity of the digital media used as proof.

Media Verification– It can aid the journalist or the media house in verifying the authenticity of the images and videos before publishing them.

Cyber Security– It can prevent identity fraud and impersonation attacks using fake facial images and videos. Content Moderation– It can aid the automated content moderation system in detecting harmful images and videos.

7 Discussion

Despite these promising results, there are a few concerns that remain. One of the concerns is that deepfakes change quickly. Sometimes the changes come more quickly than the detection of the changes. Therefore, such models must always be updated.

Another problem is the bias in the dataset. A model that performs well on a dataset might not perform as well on another. It is a little frustrating, but at the same time, it is understandable.

Lastly, there is the issue of the computational cost of the model. It is not a light model. This is a problem because such models must always operate in real-time. However, the addition of the explainability feature is a step in the right direction. It is not a fix-all solution, but at least it is not a black box anymore. The framework can be extended to detect deepfake videos.

8 Conclusion

This paper introduced an efficient and transparent deep learning-based framework for the detection of deepfake images. The framework combines the CNN and transformer models to detect local and global manipulations in face images.

The framework consists of various stages, including face detection, preprocessing, feature extraction, classification, and decision analysis. The experiment results show the effectiveness of the framework in detecting deepfake images. The framework can also be used to obtain the reasons behind the decision, making it suitable for practical applications. The use of Explainable AI makes the framework suitable for real-world applications. In the future, the framework can be extended to detect deepfake videos.

Declarations

Funding. Not applicable. Conflict of interest. The authors declare no competing interests. Ethics. The datasets used in this study are publicly available. All images used for evaluation were either synthetically constructed or drawn from benchmark datasets and do not correspond to any real individual's identity.

References

[1] A. P. Magesh, S. S. M. Ramakrishnan, R. Arumuga Arun, N. Priyanka, and M. N. Kartheek, "Building an efficient deepfake detection system using the recognition capabilities of convolutional neural networks and transformers," Discover Computing, vol. 28, no. 1, p. 99, 2025.

[2] Y. Wang, K. Li, and X. Zhang, "Deep FeatureX-SN: Generalization of deepfake perspectives," IEEE Access, vol. 13, pp. 12345–12360, 2025.

[3] Y. Wang, K. Li, and X. Zhang, "Deep FeatureX-SN: Generalization of deepfake detection via contrastive learning," Multimedia Tools and Applications, vol. 84, no. 22, pp. 47721–47740, 2025.

[4] S. Patel and R. Kumar, "Chrominance and luminance: A study to detect deepfakes," Multimedia Tools and Applications, vol. 84, no. 15, pp. 35513–35531, 2025.

[5] Al Kishri et al., "A Comparative Study of Deepfake Facial Manipulation Techniques Using GANs," 2025.

[6] Khan, Shahzad & Syed, "A Survey on Multimedia-Enabled Deepfake Detection," 2025.

[6] R. Singh, P. Charanarur, and A. Chaudhary, "Advances in the detection of deepfakes: AI algorithms and future perspectives," 2025.