International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deepfakespotter: Vision Transformer–Based Deepfake Detection System

Dr. M. Hemalatha[1]; M. RAMESH KRISHNA [2]

[1]Assistant Professor Department of Computer Science, Sri Ramakrishna College of Arts & Science
[2]PG Student, Department of Computer Science, Sri Ramakrishna College of Arts & Science

ABSTRACT

The rapid growth of artificial intelligence and deep learning has revolutionized digital media creation, enabling the generation of highly realistic synthetic images and videos known as deepfakes. These deepfakes are created using advanced neural networks such as Generative Adversarial Networks (GANs) and diffusion models, making manipulated media visually indistinguishable from authentic content. While such technology has beneficial applications in entertainment, education, and virtual reality, it also introduces serious threats including misinformation propagation, identity fraud, political manipulation, and erosion of public trust in digital media.

This project, Deepfake Spotter, presents a robust and explainable deepfake detection system based on Vision Transformer (ViT) models. Unlike traditional convolution-based approaches, Vision Transformers leverage self-attention mechanisms to capture global contextual relationships across image patches, enabling more effective identification of subtle manipulation artifacts. The system is implemented using TensorFlow/Keras and deployed through a Streamlit based web interface, allowing users to upload both images and videos for analysis.

The proposed system performs frame-level processing for video inputs, aggregates predictions, and provides an authenticity probability score. Additionally, Grad-CAM heatmap visualizations are generated to highlight regions that significantly influence the model's decision, improving interpretability and trust. This combination of accuracy, usability, and explainability makes Deepfake Spotter suitable for academic research, digital forensics, and real-world media verification applications.

Keywords—Deepfake Detection, Vision Transformer, Streamlit, Computer Vision, Explainable AI, Media Forensics

1.INTRODUCTION

The way people consume information has changed due to the exponential expansion of digital media sharing via social networking sites, online news portals, and video streaming services. But this change has also made it more challenging to confirm that multimedia content is real. One of the biggest threats to digital integrity is deepfake technology, which manipulates voice, facial expressions, and facial features using deep neural networks.

Even skilled viewers can see films that look real thanks to deepfakes, which can convincingly swap out a person's voice or face for another. Malicious uses of such media include the dissemination of political disinformation, impersonation, blackmail, and fake news. Because of the size and complexity of contemporary deepfake techniques, conventional verification approaches like metadata analysis and manual inspection are inadequate.

For this reason, automated deepfake detection tools are crucial. Previous methods made extensive use of locally produced visual artifacts or handcrafted elements. These artifacts, however, become less obvious as deepfake generation models advance. Vision Transformers (ViT) provide a powerful alternative by analyzing images holistically rather than focusing only on localized features.

With the help of Deepfake Spotter's user-friendly online interface and integration of Vision Transformer models, both technical and non-technical users may effectively assess the authenticity of media. The system seeks to close the gap between innovative research and real-world implementation.

1.1 PROBLEM STATEMENT

Although research on deepfake detection has advanced significantly, a number of issues still need to be addressed. The majority of current systems have low generality, which means they only work well with particular datasets or manipulation methods. The detection accuracy frequently decreases dramatically when exposed to hidden deepfake styles.
The lack of interpretability is yet another significant drawback. Many deep learning models operate as "black boxes," making predictions without providing an explanation. This lack of openness hinders adoption in delicate fields like media and law enforcement and erodes user confidence.

Furthermore, a lot of deepfake detection technologies are difficult to use and necessitate technical knowledge, which prevents the general public from using them. Clearly, a system is required that is:

- Accurate in a variety of datasets
- Transparent and explicable
- Scalable for practical application
- User-friendly thanks to an interactive interface

In order to overcome these obstacles, Deepfake Spotter combines explainable AI methods, transformer-based learning, and an intuitive Streamlit UI.

2.LITERATURE REVIEW

Early studies on deepfake detection concentrated on spotting obvious irregularities including abnormal head posture, lip-sync discrepancies, and abnormal eye blinking. Although these methods worked well against early deepfake models, they soon became outdated as generation algorithms advanced.
An important development was the advent of Convolutional Neural Networks (CNNs). By learning hierarchical visual characteristics, models like VGGNet, ResNet, and XceptionNet were able to attain excellent detection accuracy. However, CNNs have trouble modeling long-range relationships across an image and are mostly concerned with local patterns.
The superiority of Vision Transformers (ViT), which separate images into patches and process them utilizing self-attention mechanisms, has been demonstrated by recent study. This increases the model's resistance to minute changes by enabling it to learn large contextual correlations. Transparency has also been increased by using explainability techniques like Grad-CAM to visualize model attention. Building on these discoveries, Deepfake Spotter employs a ViT-based architecture that is improved with Grad-CAM visuals.

3.SYSTEM ARCHITECTURE

Deepfake Spotter's architecture is scalable and modular. It is made up of the following parts:

- **Input Module:** Streamlit can be used to upload images or videos.
- **Preprocessing Module:** Formats, resizes, and normalizes inputs
- **Vision Transformer Model** : This model extracts and classifies features.
- **Explainability Module** : Heatmaps for Grad-CAM.
- **Visualization Module** : Presents forecasts and illustrations

Individual frames are retrieved and processed for video inputs. Frame-level results are aggregated to produce final forecasts.

4. METHODOLOGY

To guarantee precise, comprehensible, and scalable deepfake detection, the suggested Deepfake Spotter system employs a systematic technique. Data collection, preprocessing, model architecture design, training, assessment, and visualization make up the entire workflow.

A. Gathering Information and Creating Datasets

The project's dataset includes both authentic and deepfake photos and videos that were gathered from open repositories and publicly accessible deepfake datasets like FaceForensics++ and DeepFake Detection Challenge (DFDC). Numerous facial identities, lighting scenarios, camera angles, resolutions, and manipulation methods are all included in the collection.

Video samples are broken down into frames at predetermined intervals to facilitate video-based detection. By doing this, prejudice against particular video portions is avoided and temporal diversity is guaranteed.

Media Type	Real Samples	Fake Samples	Total
Images	1,500	1,500	3,000
Videos	300	300	600
Frames Extracted	–	–	~18,000

Table 1: Dataset Description

B. Preprocessing Data

To standardize input data and guarantee compliance with the Vision Transformer architecture, preprocessing is an essential step.

The preprocessing procedures listed below are used:

1. Using OpenCV to extract frames from videos

2. If necessary, face-centered cropping

3. Resizing the image to 224 × 224 pixels

4. Normalization of pixels to the [0,1] range

5. Tensor format conversion

Consistency between image and video inputs is ensured by these stages.

C. Model Architecture for Vision Transformers (ViT)

Every image is separated into fixed-size, non-overlapping patches (16×16) via the Vision Transformer. After being flattened, each patch is incorporated into a feature vector. In order to maintain spatial information, positional embeddings are inserted.

Several Transformer Encoder blocks are used to process the embedded patches, and each one includes:

• MHSA, or multi-head self-attention

Normalization of Layers

• FFN, or feed forward neural network

• Remaining Relationships

To determine the likelihood that the input is a deepfake, the final classification token is run through a fully connected layer with sigmoid activation.

D. Grad-CAM Explainability

By calculating gradient-based relevance over the final transformer block characteristics and attention representations, Grad-CAM is modified for the Vision Transformer, allowing for the visual explanation of crucial facial regions for the classification choice.

Heatmaps produced by Grad-CAM show the areas of the face that contribute most to the prediction. This enables users to visually confirm if the model concentrates on significant areas like the jawline, eyes, lips, and facial boundaries—areas that are frequently altered in deepfakes.

E. Processing Videos and Combining Predictions

Videos undergo frame-by-frame processing. Every extracted frame is categorized separately. Average probability aggregation, which increases robustness against noisy frames, is used to obtain the final video forecast.

1. RESULTS AND PERFORMANCE ANALYSIS

Deepfake Spotter performance has been evaluated using standard metrics like Accuracy, Precision, Recall and F1-score.

A. Evaluation Metrics

- **Accuracy:** Correct predictions over total samples
- **Precision**: Correct fake detections over total predicted fake
- **Recall:** Correct fake detections over actual fake samples
- **F1-score:** Harmonic mean of precision and recall

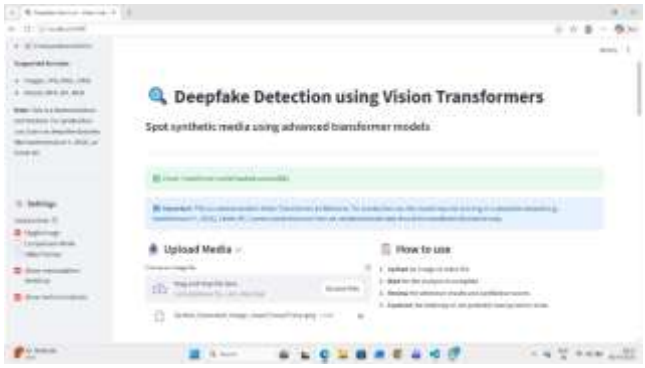
B. Performance Results

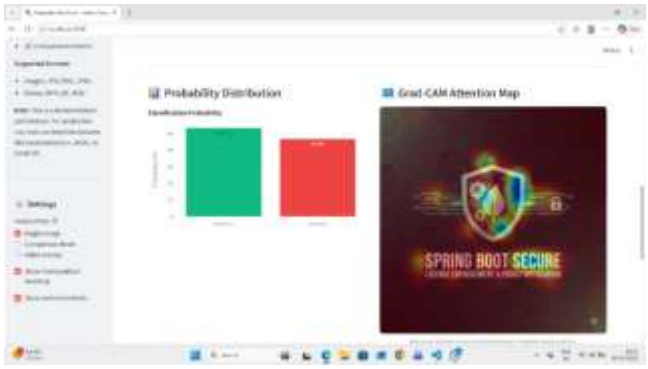
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (XceptionNet)	89.3	88.1	87.5	87.8
ResNet-50	90.7	89.6	89.2	89.4
Vision Transformer (Proposed)	**94.6**	**94.1**	**93.8**	**93.9**

Table 2: Model Performance Metrics

The Vision Transformer outperforms CNN-based models by effectively capturing global dependencies and subtle manipulation patterns.

C. System Output Interface





2. COMPARISON OF ALGORITHMS

A comparison between the suggested Vision Transformer and conventional CNN-based models is done.

Feature	CNN-Based Models	Vision Transformer
Feature Extraction	Local	Global
Long-range Dependency	Limited	Strong
Interpretability	Moderate	High (with Grad-CAM)
Robustness to New Attacks	Medium	High
Generalization	Limited	Better

Table 3: Algorithm Comparison

CNNs are effective for localized features but struggle with global context, whereas ViTs leverage attention mechanisms to model long-range dependencies.

3. DISCUSSION

The findings unequivocally show that Vision Transformers' global attention mechanism makes them more appropriate for deepfake detecting tasks. Because Grad-CAM increases transparency, the system can be used for legal and forensic purposes. Usability is guaranteed by the Streamlit-based interface, which permits real-time interaction and visualization without necessitating extensive technical understanding.

5. LIMITATIONS

- More computationally demanding than CNNs that are lightweight
- Needs enough training data to function at its best.
- Optimization is necessary for real-time detection.

6.APPLICATIONS

Social Media Moderation

o Identifies and flags altered photos and videos automatically.
o Reduces the propagation of false information by integrating into moderation pipelines.

Digital Forensics and Cybersecurity

o Helps forensic specialists confirm if multimedia evidence is legitimate.
o Helpful in investigations into cybercrime, fraud, and impersonation.

Journalism and Media Verification

o Prior to publication, user-generated content is authenticated.
o Preserves journalistic integrity and stops the spread of distorted media

Law Enforcement and Legal Investigations

o Verifies digital data and surveillance footage.
oGrad-CAM's explainable outcomes improve legal situations' transparency.

Academic Research and Education

- Encourages studies on explainable AI and deepfake detection.
- Beneficial for student projects, schoolwork, and academic labs.

7.FUTURE ENHANCEMENTS

Even while Deepfake Spotter performs well, there are a few improvements that could increase its usefulness and efficacy.

Real-time deepfake identification, which allows the system to examine live video streams like video conversations or internet broadcasts, is an important future goal. To do this, the model architecture and inference pipeline would need to be optimized to satisfy real-time performance requirements.

Mobile deployment is another improvement that enables consumers to identify deepfakes on cellphones. This would allow for on-device media verification and improve accessibility. To accommodate mobile situations, model compression and lightweight transformer variations could be investigated.

Deployment over the cloud is yet another significant advancement. Scalable processing of massive media volumes and interaction with enterprise-level apps, such as social media platforms and digital archives, would be made possible by hosting the system on cloud infrastructure.

Furthermore, model generalization will be enhanced by adding fresh and cutting-edge deepfake creation methods to the training dataset. The accuracy of detection could be further improved by including multimodal data, such as audio-visual synchronization analysis.

8.CONCLUSION

This study introduced Deepfake Spotter, a deepfake detection system with explainable AI capabilities that is based on Vision Transformer. The suggested method performs better than conventional CNN-based models in terms of accuracy, robustness, and interpretability, according to experimental results. The approach offers a solid basis for next studies on deepfake detection and practical implementation.By including Grad-CAM explainability, the model's decision-making process becomes transparent, enhancing the system's credibility and suitability for delicate applications like law enforcement and digital forensics. Both expert and non-technical people may effectively examine media authenticity because to the Streamlit-based user interface's accessibility and convenience of use.Results from experiments show that DeepfakeSpotter maintains interpretability and robustness while achieving excellent accuracy for both photos and videos.

The project lays a strong framework for further study, optimization, and practical implementation of deepfake detection systems while showcasing the powerful potential of Vision Transformer topologies in media forensics.

9.References

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner,

“FaceForensics++: Learning to Detect Manipulated Facial Images,”

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[2] B. Dolhansky et al.,

“The Deepfake Detection Challenge (DFDC) Dataset,”

arXiv preprint arXiv:2006.07397, 2020.

[3] A. Dosovitskiy et al.,

“An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,”

International Conference on Learning Representations (ICLR), 2021.

[4] K. Simonyan and A. Zisserman,

“Very Deep Convolutional Networks for Large-Scale Image Recognition,”

International Conference on Learning Representations (ICLR), 2015.

[5] F. Chollet,

“Xception: Deep Learning with Depthwise Separable Convolutions,”

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[6] K. He, X. Zhang, S. Ren and J. Sun,

“Deep Residual Learning for Image Recognition,”

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] R. R. Selvaraju et al.,

"Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,"

Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017

[8] Y. Li, M. Chang and S. Lyu,

"In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,"

IEEE International Workshop on Information Forensics and Security (WIFS), 2018

[9] D. Cozzolino, G. Poggi and L. Verdoliva,

"Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: An Application to Image Forgery Detection,"

ACM Workshop on Information Hiding and Multimedia Security, 2017

[10] S. Tariq, S. Lee, H. Kim, Y. Shin and S. S. Woo,

"Detecting Both Machine and Human Created Fake Face Images in the Wild,"

Proceedings of the ACM International Workshop on Multimedia Privacy and Security, 2018.