Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

Data-Efficient Transformer Architectures for Image-Level Facial Forgery Detection: A Comparative Evaluation of ViT and DeiT

Akshatha G*

*Corresponding author, Assistant Professor, Department of Computer Science& Engineering, B.M.S College of Engineering, Affiliated to Visvesvaraya Technological University, Belagavi, India. E-mail: akshu965@gmail.com

Kempanna M

Associate Professor, Department of Artificial Intelligence and Machine Learning, Bangalore Institute of Technology, Visvesvaraya Technological University, Bangalore, India. E-mail: kempsindia@gmail.com

Preethi Kolluru Ramanaiah

Cloud Architect, Lead of AI initiative Program, Ernst & Young LLP, New York, USA. E-mail: preethiram4@gmail.com

Bandi Doss

Department of ECE, CMR Technical Campus, Hyderabad, Telangana, India. E-mail: dasalways4u@gmail.com

Ramani S

Professor, ECE Department, Sreenidhi Institute of Science and Technology, India. E-mail: dr.ramani2017@gmail.com

Kirubakaran Rangasamy

Assistant Professor, Department of Biotechnology, Vinayaka Mission`s Kirupananda Variyar Engineering College, Salem (Vinayaka Mission`s Research Foundation). India. E-mail: kirubakaran@vmkvec.edu.in

Journal of Information Technology Management, 2025, Vol. 18, Issue 1, pp. 17-33
Published by the University of Tehran, College of Management
doi: https://doi.org/10.22059/jitm.2026.106252
Article Type: Research Paper

Received: August 16, 2025
Received in revised form: September 25, 2025
Accepted: December 27, 2025
Published online: January 20, 2026

Abstract

The rapid development of deepfake technologies has increased the demand for a credible and inter-pretable system for facial forgery detection. This study compares two transformer-based architec-tures—Vision Transformer (ViT) and Distilled Data-Efficient Image Transformer (DeiT)—for de-tecting real and manipulated facial images. The study aims to measure performance in terms of de-tection as well as interpretability and to address the weaknesses of traditional convolutional models. Data augmentation was applied, and a balanced dataset containing 8,000 real and fake images was constructed; both models were then fine-tuned under the same training environment. The explanatory ability of the models was incorporated using LIME. Experimental findings indicate that both models perform well, with DeiT being slightly more accurate at 94.62% than ViT at 93.6%, alongside faster convergence rates and less overfitting. Visualization of the focus on important facial areas confirms that the models reliably register synthetic artifacts. Although promising, generalization across dif-ferent datasets and enhancement of real-time performance remain challenges. Overall, the results validate transformer architectures—especially DeiT—as powerful and explainable deepfake detec-tion algorithms, valuable for ensuring safe and transparent digital media forensics.

Keywords: Vision Transformer; Data-Efficient Image Transformer; Face Forgery Detection; Ex-plainable AI; Transformer Models

Introduction

Advanced development of artificial intelligence (AI) has changed the way images are generated, making it possible to create hyper-realistic facial forgeries known as deepfakes. These artificial images pose serious risks to digital media integrity, security, and ethics, as they can be used to misinform, steal identities, and manipulate individuals. Ensuring the authenticity of visual content has therefore become a critical challenge in the modern digital landscape, requiring the development of effective and interpretable detection systems (Korshunov and Marcel, 2018). Early forgery-detection research largely relied on Convolutional Neural Networks (CNNs) and their variants to distinguish real from synthetic faces. Despite strong performance in controlled scenarios, their accuracy deteriorated significantly when they encountered unseen manipulations. CNN-based detectors were limited by poor generalization, reduced interpretability, and an inability to model long-range dependencies within images, restricting their effectiveness in complex deepfake contexts (Yasser et al., 2023).

Transformer-based architectures have emerged as a promising solution to overcome these shortcomings. Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), replacing convolutional operations with self-attention to capture global relationships among image patches. Building on this, Touvron et al. (2021) proposed the Data-Efficient Image

Transformer (DeiT), which employed knowledge distillation to achieve high accuracy even with limited training data. These advancements provide a pathway for developing efficient and scalable facial-forgery detection systems. In this work, transformer-based models—specifically ViT-Base and DeiT-Base (distilled)—are applied to real versus fake facial image classification. To enhance transparency and trustworthiness, their reasoning processes are visualized and interpreted using Explainable Artificial Intelligence (XAI) techniques, particularly the Local Interpretable Model-agnostic Explanations (LIME) framework

In general, the study will yield a clear, interpretable, and data-efficient transformer-based framework for facial forgery detection. The paper is one of the contributions to the emerging explainable deep learning in digital forensics as it integrates attention-based visual perception with interpretable analysis, improving the reliability and ethical responsibility of AI-based image verification.

Literature Review

Detection of forgery in faces has become a central research focus with the rise of deep learning–based generative models. Altaei (2022) demonstrated that CNNs can effectively extract texture cues, although their performance drops significantly when evaluated on unfamiliar datasets. Dolhansky et al. (2019) addressed dataset limitations by releasing the Deepfake Detection Challenge (DFDC), a large-scale benchmark for evaluating models in realistic conditions; however, DFDC results also revealed that early CNN-based detectors exhibited poor generalization to unseen manipulations. Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), replacing convolutions with self-attention to capture long-range dependencies, achieving stronger performance but requiring large datasets—an obstacle for forensic applications.

Deng et al. (2022) built an EfficientNet-V2–based detector that achieved promising results but suffered from dataset dependency, while James et al. (2025) proposed X-FACTS, an explainable CNN framework supporting improved interpretability, though still reliant on extensive annotated data. Kumar and Selvam (2025) further advanced hybrid modeling by integrating convolutional layers with residual attention modules, improving accuracy but again requiring large, diverse datasets. Korshunov and Marcel (2018) were among the earliest contributors, using the VidTIMIT dataset to identify visual artifacts as markers of deepfake manipulation and demonstrating the importance of dataset diversity for model reliability.

Nagahisarchoghaei et al. (2023) emphasized explainability as a growing priority in AI-driven forensics, driving interest in more transparent detection frameworks. Omodunbi et al. (2025) designed a multilayer CNN achieving 95% accuracy on Celeb-DF, though its performance decreased on unfamiliar data. Omotosho et al. (2021) developed a CNN for real-time face recognition, showing that hierarchical spatial learning enhances verification but fails

under high noise. Oulad-Kaddour et al. (2023) generated artificial data to strengthen gender classification robustness, illustrating how synthetic augmentation can improve model stability in tasks like deepfake detection. Pai and Sharmila (2023) applied CNN-based segmentation for face-mask detection, demonstrating competitive performance even with limited feature depth.

Rajagukguk et al. (2024) implemented EfficientNet models to improve both accuracy and computational efficiency, though the models remained challenging to interpret. Touvron et al. (2021) proposed DeiT, which used knowledge distillation to attain high accuracy with far smaller training sets, making it ideal for deepfake detection, where dataset diversity is often constrained. Yasser et al. (2023) evaluated DenseNet and Xception classifiers, which performed well on existing datasets but struggled with new or unseen fake formats. Zhou and Yu (2022) developed an attention-based CNN focusing on discriminative regions such as the eyes and lips, demonstrating that attention mechanisms enhance detection accuracy and paving the way for modern transformer-based approaches.

Across these works, Table 1 summarizes the key ideas, contributions, datasets, metrics, and limitations. Early CNN-based approaches were susceptible to overfitting, lacked interpretability, and displayed weak cross-domain generalization. Transfer-learning models offered incremental improvements but remained reliant on localized features. More recent efforts highlight explainability and robustness as essential components for trustworthy forensic systems, motivating the migration from traditional CNNs to transformer-based architectures such as ViT and DeiT.

Table 1. Comparative Analysis of Prior Works in Facial Forgery Detection

Studies	Method	Dataset	Accuracy / Metric	Key Contribution & Limitation
Dolhansky et al. (2019)	CNN-based (TamperNet, XceptionNet variants)	DFDC Preview Dataset (5,214 videos)	Precision: 0.930; Recall: 0.268	Provided a large, ethically sourced dataset with improved metrics but revealed poor generalization on unseen manipulations.
Korshunov & Marcel (2018)	GAN-based Deepfake generation; VGG & Facenet testing; IQM+SVM detection	VidTIMIT (Deepfake video dataset)	FAR: 85.62% (VGG); 95.00% (Facenet); EER: 8.97%	Introduced the first public deepfake dataset and baseline metrics; detection models showed high false acceptance and weak robustness.
Kumar & Selvam (2025)	R-CNN with Gabor + MobileNetV2/VGG16	Mixed real & GAN-generated faces	85% vs. 70% (GAN-based)	Developed a robust hybrid CNN model that generalized well across domains but lacked deeper architecture-level explainability.
Mansoor & Iliev (2025)	CNN + Network Dissection (ResNet50, VGG16, InceptionV3)	CelebAMask-HQ	Acc: up to 86.2%; IoU: 0.7601	Applied visual interpretability to CNN decisions, yet analysis

				remained confined to convolutional architectures without transformer adaptation.
Nagahisarchoghaei et al. (2023)	Survey on Explainable AI (LIME, SHAP, interpretable models)	Case study: BBC News	51% confidence (Tech class example)	Offered an extensive overview of explainable AI methods, underscoring the need for transparent decision-making in deepfake detection.
Nida et al. (2021)	ELA + CNN (VGG, Inception, ResNet)	Real & Fake Face Detection (Yonsei Univ.)	91.97% train; 64.49% test (VGG-16)	Demonstrated that compression artifact detection aids fake identification but suffered from poor generalization across manipulations.
Omodunbi et al. (2025)	Transfer learning (EfficientNet, Xception, VGG19)	Kaggle (140,000 images)	94% (EfficientNet)	Demonstrated strong performance through preprocessing and transfer learning, but lacked real-time applicability and interpretability.
Rajagukguk et al. (2024)	ResNet50 CNN	589 real, 700 fake faces	Train: 76.07%; Test: 53%	Demonstrated the overfitting issue in small datasets, highlighting the need for better generalization methods.
Yasser et al. (2023)	EfficientNet-B4 vs. XceptionNet	FF++, Celeb-DF(v2)	AUC: 95.59% (FF++); 98.75% (Celeb-DF v2)	Compared CNN-based architectures, showing that robust models still struggle with evolving forgery patterns and dataset bias.

Akshatha and Kempanna (2025) surveyed deep learning approaches for fake-face identification, emphasizing ensemble methods and the necessity for adaptive, real-time detection. Arshed et al. (2023) fine-tuned ViT on DFDC, improving discrimination and providing interpretable attention heatmaps. Gong and Li (2024) further demonstrated that attention-based interpretability modules increase transparency and model stability, while Lad (2024) introduced a human-centered XAI framework integrating visual attention with LIME explanations, highlighting the growing demand for accountability in detection systems. Mansoor and Iliev (2025) extended this focus on explainability by applying LIME and SHAP to improve transparency in forensic models.

Nida et al. (2021) reported persistent overfitting in CNN-based deepfake detectors—a challenge also acknowledged in earlier studies by Dolhansky et al. (2019) and Korshunov and Marcel (2018). Rahman et al. (2023) contributed to forgery localization by using encoder–decoder networks for facial segmentation, demonstrating how spatial coherence can support manipulation detection. Omodunbi et al. (2025) reaffirmed the value of interpretability through their multilayer CNN design, and Zhou and Yu (2022) showed that attention mechanisms focusing on key facial regions enhance detection robustness. Together, these works align closely with the principles underlying transformer-based forensic paradigms.

The literature, for the most part, points to a definite shift from CNNs, which extract local texture cues, to transformers, which model global dependencies and provide explainable outputs. Although advancement has been made, a good number of detectors are still finding it hard to scale up and be transparent. The use of ViT and DeiT is thus a natural progression as both are able to effectively learn context-rich representations and can be easily integrated with XAI tools like LIME to facilitate transparent, trustworthy, and socially acceptable facial forgery detection.

Methodology

The methodological framework of this research was crafted to allow it to be repeated, make it clear, and show that it is an efficient way to distinguish between fake and real face images with the help of transformers. The procedures incorporate dataset preparation, augmentation, data splitting, model development, training, explainability integration, and evaluation, each being organized in such a way that replication under the same conditions ispossible.

Dataset Preparation

The study relied on two publicly accessible datasets: The Real and Fake Face Detection (CIPL Lab, Kaggle) dataset with 1081 authentic and 960 counterfeit images, and the FaceForensics (Kaggle) dataset with 650 sophisticated forgery images. As a result of these datasets, the researchers had various examples of facial manipulation, going from simple edits to complex deepfake synthesis. The dataset was balanced and diversified through augmentation, which was then finalized in a corpus of 8000 images, 4000 real, 4000 fake. This made sure there was enough data volume, class balance, and variability for the models to be trained and evaluated in a robust way.

Data Augmentation

Augmentation was performed using TensorFlow and Keras libraries to reduce overfitting and better reflect real-world variability. The changes consisted of random rotations, width and height shifts, zoom greater than or equal to 10 %, and horizontal flips. Vertical flips were only done in some cases to increase invariance, and the fill mode was set to “constant” to keep the boundary level. Every change kept the semantic side of things, so the labels were still correct, and real faces remained real. These methods have led to better model generalization and less dataset bias, which has been a frequent problem in deepfake detection (Rajagukguk et al., 2024).

Data Splitting

After augmentation, the data splits for training 80%, validation 10%, and testing 10% were in line with the advice of Akshatha et al. (2026).

- Training set (6400 images) – the model uses this data for learning and updating the parameters.
- Validation set (800 images) – monitored to prevent overfitting and guide hyperparameter tuning.
- Testing set (800 images) – kept for a fair performance evaluation without any bias.

This stratified split ensured balanced representation and reliable statistical assessment across all stages.

Experimental Workflow

It was made to realize the implementations of two transformer-based models, Vision Transformer and Distilled-based DeiT, by means of PyTorch and Hugging Face Transformers in Google Colab with the use of the GPU accelerator (NVIDIA Tesla T4, 16 GB VRAM). In both models, to carry out binary classification, their default classification head was replaced with a two-logit output layer. The operations included data preprocessing, fine-tuning, XAI integration, and metric evaluation, which were carried out under the same computational conditions for a fair comparison as presented in Figure 1.

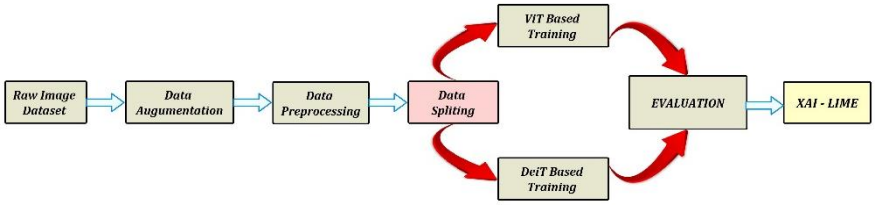

Figure 1. Proposed System

Vision Transformer (ViT-Base) Model

The ViT-based architecture (Dosovitskiy et al., 2020) is a major change in the way features are extracted. The model moves from the use of convolutions to self-attention. In fact, Figure 2 shows how ViT splits each 224 × 224 face image into 16 × 16 patches without overlapping; thus, the number of patches is 196. A patch is flattened, and then a linear projection is done to get a 768-dimensional embedding. Positional encodings are added to keep the spatial order since transformers are not inherently spatially aware.

The encoded patch sequence is then processed through 12 transformer encoder layers, each composed of:

- Multi-Head Self-Attention (12 heads): Learns long-range dependencies across all patches.
- Feed-Forward Layers (MLP): Expands the feature dimension to 3072 before projecting back to 768.
- Residual Connections and Layer Normalization: Facilitate gradient flow and training stability.

A [CLS] token, which is added before the patch sequence, represents the global image. The embeddings of this token after the last layer are fed to a fully connected classifier to predict real or fake labels. The architecture of ViT is very powerful in detecting faint global inconsistencies of lighting, symmetry, or texture, helping the model to recognize those patterns that are most likely to be ignored by convolutional filters.

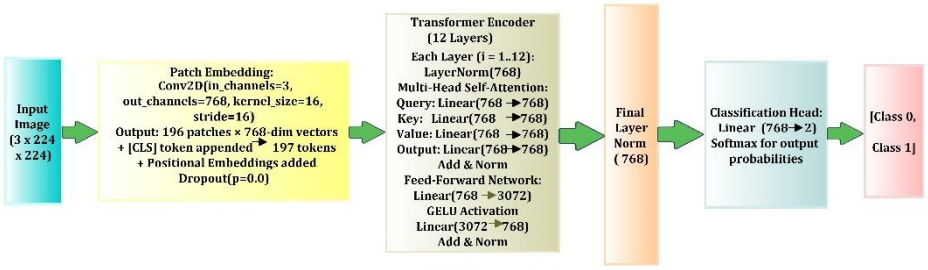

Figure 2. ViT architecture diagram

DeiT-Base Distilled Model

Data-Efficient Image Transformer (DeiT-Base Distilled) by Touvron et al. (2021) changes the ViT's design to make it more data-efficient by means of knowledge distillation. The DeiT architecture, depicted in Figure 3, is similar to ViT in the sense that it uses patch embedding but also has a new distillation token ([DIST]) that it introduces. This token communicates with a teacher network, usually a convolutional backbone like RegNet, thus the transformer can get the inductive biases that the CNNs have learned.

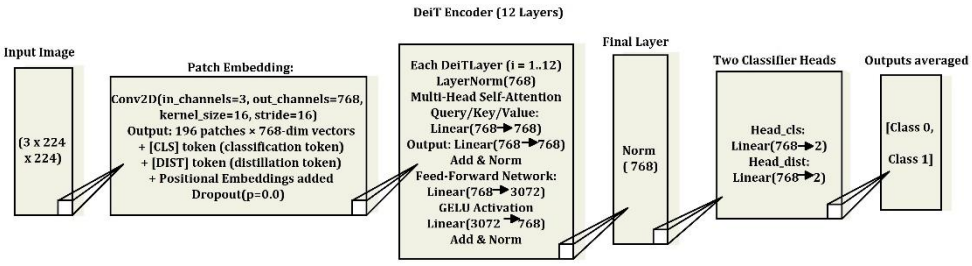

Figure 3. DeiT architecture diagram

Key architectural features include input size: 224 × 224; Patch size: 16 × 16 (196 patches). Embedding dimension: 768; MLP hidden layer size: 3072. Sequence length: 198 tokens (196 patches + [CLS] + [DIST]). 12 transformer encoder layers with 12 attention heads. DeiT's twin-token strategy gives rise to the possibility that the predictions might be made either from the [CLS] token, the [DIST] token, or a weighted mixture of both. Such a hybrid system facilitates feature transfer and helps the model to generalize better when the size of the dataset is moderate, thus it mitigates the problem of ViT being inherently heavily reliant on large-scale data. The DeiT architecture, hence, represents a compact and readable method of detecting fakes, especially when the availability of data is limited.

Architectural Comparison of ViT-Base and DeiT-Base Distilled

Table 2 . Architectural Comparison between ViT and DeiT-Base Distilled Models

Feature	ViT-Base (google/vit-base-patch16-224)	DeiT-Base Distilled (facebook/deit-base-distilled-patch16-224)
Model Family	ViT	DeiT (Data-efficient Image Transformer)
Distillation Token	No	Yes (adds an extra [DIST] token)
Input Image Size	224 × 224	224 × 224
Patch Size	16 × 16	16 × 16
Patches per Image	14 × 14 = 196	14 × 14 = 196
Special Tokens	1 ([CLS])	2 ([CLS] + [DIST])
Total Sequence Length (Tokens)	197	198
Transformer Encoder Layers	12	12
Hidden Size	768	768
Attention Heads	12	12
MLP Hidden Layer Size	3072	3072
Parameters	~86M	~86M
Output Head	Classification logits from [CLS] token	Classification logits from [DIST] token (or averaged)

Table 2 compares both transformer architectures used in this study. DeiT maintains ViT's architecture but achieves higher data efficiency via knowledge distillation, balancing transformer-level context modeling with CNN-like locality awareness.

Training Setup

For proper comparison of their capacities, both models were fine-tuned under the same hyperparameter configurations to avoid any bias and to make sure that any difference in the performance is due to the different architectural designs only. Each image fed to the network was changed in size to 224×224 pixels, and normalized to the range of [–1, +1]. Learning and convergence of the model were made by taking a batch size of 32 and 30 training epochs. To improve the classification accuracy, the Adam optimizer, with a learning rate of 2×10^{-5}, was used together with the Binary Cross-Entropy loss function. To speed up convergence and to leverage the prior feature representations, both models were set up with pretrained ImageNet weights from the very beginning. In order to avoid overfitting, the process of early stopping based on validation loss was also used. This setting, which is in line with the method as per Touvron et al. (2021), was the reason for steady convergence and the possibility of repeating the experiment for both transformer architectures.

Explainable AI (XAI) Integration

LIME was used in both models to make their working clear to the user. LIME changes the different parts of the image and notices the changes in the prediction to find the features that influence the most. The heatmaps produced in Figure 8 and Figure 9 show the areas like eyes, cheeks, and lips, which were the region's most discriminative in fake-face detection. The study here confirmed that both ViT and DeiT models concentrate on the correct parts of the faces, which are the features for the background noise, thus they are transparent and their

decisions can be trusted. The use of LIME is in line with the previous work that was focused on explainable AI for digital forensics (Mansoor & Iliev, 2025), which is a way of making the system's decisions understandable and ethically deployable.

Evaluation Metrics

The model's performance was assessed through various important metrics that provide a complete picture of the model's accuracy and reliability. Accuracy was a measure of the total correctness of the classification, while precision was used to show the model's capacity to reduce false positives to a minimum. Recall or sensitivity was the measure of how the model was able to find the largest number of true positive cases, and the F1-score was a balanced measure of precision and recall. Moreover, the confusion matrix gave a detailed representation of the error distribution in different classes. All these metrics together allowed for a comprehensive evaluation of the detection's robustness, which involved not only the predictive accuracy but also the reduction of false alarms to a minimum, thus making deepfake detection systems practically usable and trustworthy.

Results

Outcomes of this research deliver an in-depth comparison of the ViT and DeiT models for identifying Fake vs. Real images with augmented datasets. Both models' predictive performance and learning efficiency were analyzed, compared, and judged through the evaluation metrics, classification reports, confusion matrices, and training-validation trends.

Table 3. Classification Report of ViT and DeiT Models for Fake vs Real Image Detection

Model	Class	Precision (%)	Recall (%)	F1-Score (%)	Support (Samples)	Overall Accuracy (%)	Macro Avg (%)	Weighted Avg (%)
ViT	Fake	95	92	93	391	93.62	94	94
	Real	93	95	94	409			
DeiT	Fake	93	96	95	402	94.62	95	95
	Real	96	93	95	398			

Table 3 represents the classification report that compares the performance of ViT and DeiT models on the Fake vs Real image detection task. The ViT model has an overall accuracy of 93.62%, with macro and weighted averages of 94%, which means that the classification of Fake and Real categories was strong and balanced. On the other hand, the DeiT model showed a slightly higher accuracy of 94.6 %, with both macro and weighted averages of 95% remaining stable, which indicates better generalization and more refined feature extraction capabilities. Also, both models have made it to very high F1-scores over 93%; thus, a minimum difference between precision and recall has been achieved, and it is a manifestation of the excellent and dependable detection of image authenticity throughout the dataset.

The confusion matrices of the ViT and DeiT models offer a profound understanding of their classification capability for the normalized setting of the Fake and Real images distinction. As per Figure 4, the confusion matrix for the ViT model shows that the system correctly recognized 361 Fakes out of 391 and wrongly labeled 30 as Reals. Similarly, out of 409 Real images, the algorithm confirmed 388 as Real, but 21 were falsely predicted as Fake. This allocation brings about the sum of correct predictions to 749 out of 800 test samples, which is equivalent to the overall accuracy of 93.62 percentage. The comparatively low count of the wrongly classified items, 51 in total, signifies that the model is a strong one in discriminating between the fake and real images even when they are visually similar.

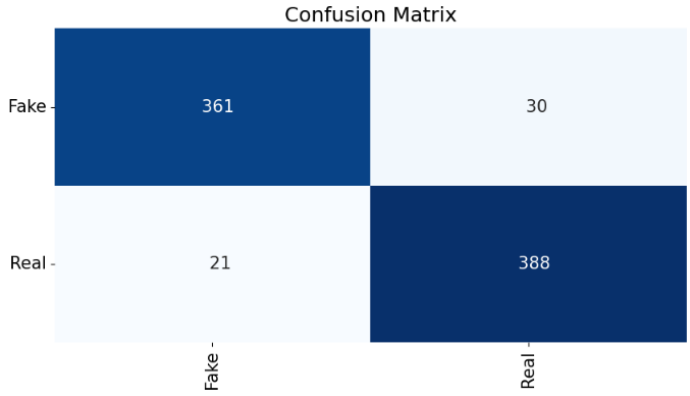

Figure 4. Confusion Matrix of ViT Model for Fake vs Real Image Detection

On the other hand, the DeiT model shown in Figure 5 had a slightly better performance, with 386 Fake instances being correctly classified as Fake (true positives) and 371 Real instances being accurately predicted as Real (true negatives). The model made a mistake in classification by detecting 16 Fake images as Real (false negatives) and 27 Real images as Fake (false positives). From 800 total instances, the model made 757 correct predictions, resulting in an overall accuracy of 94.62%. The increased true positive and true negative counts, together with fewer false negatives, signify that the DeiT model is not only as precise as the ViT model but also has a higher recall of Fake image identification. In essence, both models are effective, and DeiT has a slight edge in identifying Fake and Real images correctly. The confusion matrix assessments corroborate that DeiT has better generalization and discriminative capability, which is consistent with its higher overall accuracy and balanced classification metrics reported in Table 1.

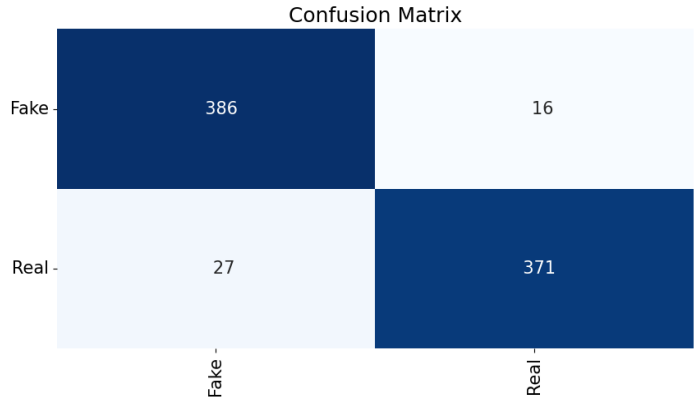

Figure 5. Confusion Matrix of DeiT for Fake vs Real Image Detection

The performance of both models over 30 epochs has been analyzed by us to understand their learning stability, convergence, and generalization. From the charts of Figure 6, we can see that the ViT model was able to continuously enhance its performance in terms of both accuracy and loss. Training loss went down very quickly from about 46 percentage at the first epoch to close to 10% at epoch 18, with a little bit of variation of 11%and 17% happening in the last part of the training, thus converging stably with a minimum of overfitting. Similarly, validation loss dropped from 32% to around 22% by epoch 5 and then stayed within the 20 to 25% band for the rest of the time. In the same way, training accuracy was increased from 77% to almost 98%, whereas validation accuracy was improved from 87% to a stable 92 to 94%, thus confirming that the model generalizes well. On the other hand, the DeiT model Figure 7 had better optimization and performance stability as well. Its training loss was near 45% at the beginning and kept on going down to around 5%by epoch 30, while validation loss was reduced from 43%to nearly 8%, thus showing very little difference between the two curves, an indicator of strong generalization and less overfitting. Training accuracy went up very fast from 78% to almost 100%, and validation accuracy increased from 76% to about 95% in the last epochs.

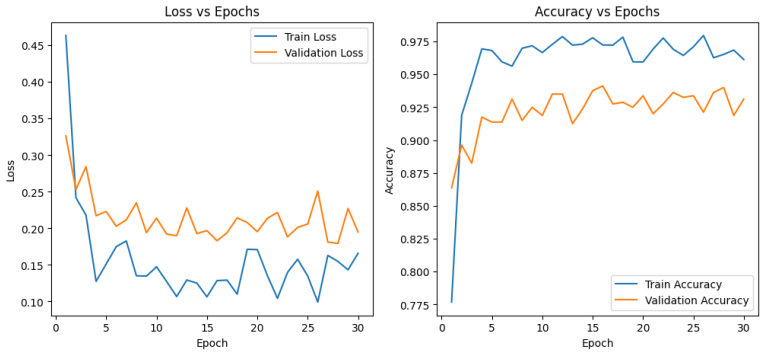

Figure 6. Training and Validation Loss and Accuracy Trends of ViT Model

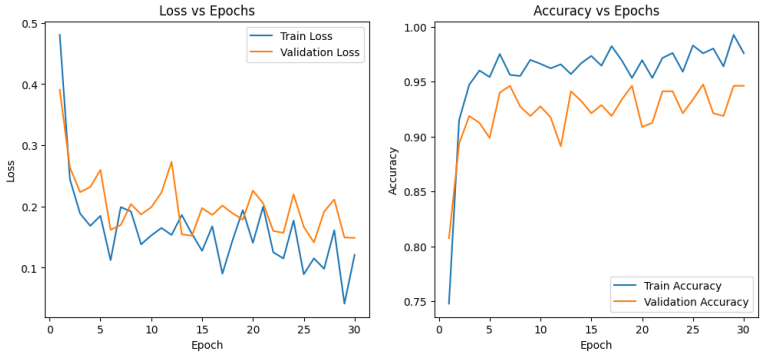

Figure 7. Training and Validation Loss and Accuracy Trends of the DeiT Model

By comparison, the DeiT model shows a lower final validation loss of 8%, compared to 20–25%, and a better final validation accuracy of 95% versus 94% for the ViT model. These results indicate better convergence behavior, higher optimization efficiency, and a greater ability to both extract and generalize discriminative features. The steady decrease of both training and validation losses, along with the stability of the accuracy curves are evidence that DeiT has more effective learning dynamics for Fake vs. Real image detection.

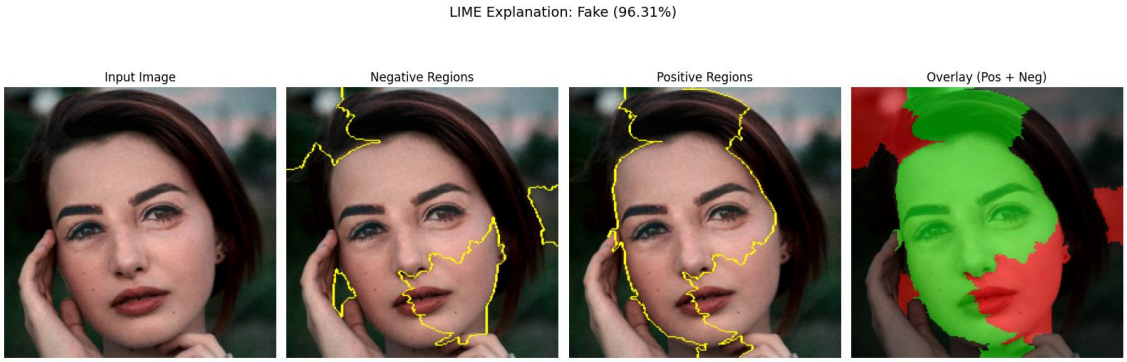

Figure 8. Lime output for ViT

The ViT model shown in Figure 8 was the one that recorded the highest confidence in classification of 96.31%. It was thus able to recognize the analyzed image as Fake in the strongest way. Such a high confidence level can be seen as a clear indication of the model's powerful capacity to differentiate real and artificially generated facial textures. LIME was the method that was used to find the local approximation for ViT's decision by means of perturbing the input image and looking into the changes in prediction resulting from the perturbations. The LIME visualization depicts those parts of the image that helped most, which is highlighted in green in the classification of Fake lying on the middle facial features like cheeks, nose bridge, and lower jawline. Faces that obviously played the most crucial role in labelling the image as Fake. There were also negative regions in red around the hairline and background, which, however, had a very little or no influence at all, or an opposite one. The findings here demonstrate that the ViT model is very effective in directing its attention to those facial features which are perceptually most relevant as well as to the tiniest of the

artifacts that are typical of the manipulated images and that were left there without any doubt as a result of the editing, thus not only verifying its predictive accuracy but also the representational power of its XAI visualization.

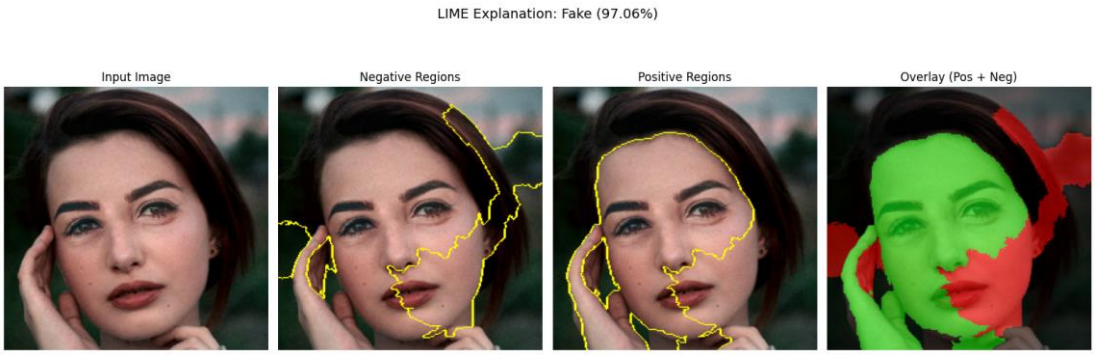

Figure 9. Lime output for DeiT-ViT

Similarly, the DeiT model, shown in Figure 9, classified the same image as Fake with a confidence level of 97.06%, reflecting superior discriminative capability. Using the same LIME explainability method, the DeiT visualization revealed that the most influential regions, shown in green, were concentrated around the forehead, cheeks, nose, and lips—key zones for texture irregularities and subtle distortions. In contrast, negative regions red along the hairline, fingers, and background contributed minimally to the final prediction. These visual cues demonstrate that DeiT effectively localizes and interprets generative inconsistencies while maintaining transparent decision reasoning. The LIME-based interpretability confirms that both transformer models not only achieve high detection accuracy but also provide clear, human-understandable explanations, reinforcing their reliability for ethical and explainable deepfake detection.

Discussion

The results of the given research indicate that both the ViT and DeiT models have been successfully utilized in overcoming the task of identifying fake and real facial images, thus achieving the goal of building a solid and understandable framework to identify authenticity in images. Both models worked well, although DeiT was a bit better than ViT with a total accuracy of 94.62% as opposed to 93.62%. This gain is an indication of increased data efficiency and better efficiency in the attention mechanisms of DeiT, which facilitates faster and more efficient learning. The Precision, Recall, and F1-scores of both models were more than 93%, indicating high generalization on real and fake classes. The F1-score of 95% of the DeiT shows that it is more consistent in identifying small generative artefacts that are vital in the detection of deepfakes. This is also confirmed by the confusion matrix analysis, which indicates that DeiT incorrectly recognized 43 images as compared to 51 by ViT.

Also, the smaller validation gap around 3%of DeiT compared to ViT around 8 to10% demonstrates better convergence and minimized overfitting. The incorporation of LIME

offered valid information on the decision-making process of the two models. The visual interpretations revealed that ViT and DeiT paid more attention to main facial areas, including cheeks, nose, and lips, that have minute texture variations reflecting manipulation. The attention maps of DeiT were coherent and more local, thus assuring the presence of stronger focus and interpretability. It proves that DeiT is not only functional but also renders clear and reliable forecasts, which is crucial in the field of forensics. DeiT was also more data-efficient and stable, converged much faster, and had a validation loss near 8% in comparison to ViT with a range of 20 to 25%. This efficiency renders DeiT more applicable in real world application where there is a scarcity of resources.

On the whole, these findings correspond to the existing literature that transformer-based architectures have superiority over CNNs in forgery detection because of their capacity to capture global dependencies. The current work builds on this knowledge and demonstrates that DeiT has the same strengths as ViT with fewer data and computation requirements. To sum up, both ViT and DeiT show good results in real and fake face recognition. Nevertheless, DeiT has a superior accuracy, interpretability, and efficiency ratio, which proves its feasibility as a real-life and transparent approach to digital forensics and media authenticity verification relying on AI.

Conclusion

This paper examined the efficacy of transformer-based architectures in the classification of facial forgery. Both models showed a high ability to differentiate between original and fabricated faces; nevertheless, the DeiT model showed slightly better results, with higher accuracy and generalization with lower overfitting. The incorporation of LIME-based interpretability further asserted that all the models were always keen on key parts of the face features, like the cheeks, lips, and nose, that could carry out clear and reliable model decisions. These results emphasize the possibility of distilled data-efficient transformer models as effective and reliable methods of digital media forensics in an explainable way. The findings imply that transformer architectures and especially DeiT can be used as scalable backbones with applications in the real-world of deepfake detection, where transparency and computational efficiency are critical requirements. The future studies may involve cross-dataset validation and real-time implementation. On the whole, this work supports the role of interpretable AI as a tool for enhancing digital trust and addressing the threat of synthetic media that is emerging.

Acknowledgements

The authors sincerely thank B.M.S. College of Engineering, Bengaluru, and Bangalore Institute of Technology, Bengaluru, for their continuous support and encouragement in

carrying out this research. Their institutional resources and academic environment greatly contributed to the successful completion of this work.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Akshatha G, Kempanna, M., Ashoka, S. B., & Kunta, J. P. K. C. (2026). Deep Learning for Facial Forgery Detection: Performance Evaluation of DenseNet201, InceptionV3 and ConvNeXt.

Akshatha, G., & Kempanna, M. (2025). Review of deep learning strategies in modern fakeface identification systems. *Grenze International Journal of Engineering and Technology*, *11*(2), 4964–4974.

Altaei, M. S. M. (2022). A detection of deep fake in face images using deep learning. *Wasit Journal of Computer and Mathematics Science*, *1(4)*, 60-71.

Arshed, M. A., Alwadain, A., Faizan Ali, R., Mumtaz, S., Ibrahim, M., & Muneer, A. (2023). Unmasking deception: empowering deepfake detection with vision transformer network. *Mathematics*, *11*(17), 3710.

Deng, L., Suo, H., & Li, D. (2022). Deepfake Video Detection Based on EfficientNet-V2 Network. *Computational Intelligence and Neuroscience*, 2022(1), 3441549.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: datasets, algorithms, and challenges. *Electronics*, *13*(3), 585.

James, U. U., Olarinoye, H. S., Uchenna, I. R., Idika, C. N., Ngene, O. J., Ijiga, O. M., & Itemuagbor, K. (2025). Combating Deepfake Threats Using X-FACTS Explainable CNN Framework for Enhanced Detection and Cybersecurity Resilience.

Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Kumar, M. and Selvam, A., 2025. Deep Fake Face Detection Using Advanced R-CNN Architectures. *IJSAT-International Journal on Science and Technology*, *16*(2).

Lad, S. (2024). Applied Ethical and Explainable AI in Adversarial Deepfake Detection: From Theory to Real-World Systems. *Journal of Artificial Intelligence General science (JAIGS)*, *6*(1), 126-137.

Mansoor, N., & Iliev, A. I. (2025). Explainable AI for deepfake detection. *Applied Sciences*, *15*(2), 725.

Nagahisarchoghaei, M., Nur, N., Cummins, L., Nur, N., Karimi, M. M., Nandanwar, S., ... & Rahimi, S. (2023). An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics*, *12*(5), 1092.

Nida, N., Irtaza, A., & Ilyas, N. (2021). Forged face detection using ELA and deep learning techniques. In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, 271-275.

Omodunbi, B. A., Sobowale, A., & Soladoye, A. Detection of Image-Based Deepfake using Deep Transfer Learning Algorithms.

Omotosho, L. O., Ogundoyin, I. K., Oyeniyi, J. O., & Oyeniran, O. A. (2021). A real time face recognition system using Alexnet deep convolutional network transfer learning model. *Journal of Engineering Studies and Research*, *27*(2), 82-88.

Oulad-Kaddour, M., Haddadou, H., Vilda, C. C., Palacios-Alonso, D., Benatchba, K., & Cabello, E. (2023). Deep learning-based gender classification by training with fake data. *IEEE Access*, *11*, 120766-120779.

Pai, G., & Sharmila, K. M. (2023). Semi-Dense U-Net: A Novel U-Net Architecture for Face Detection. *International Journal of Advanced Computer Science and Applications*, *14*(6).

Rahman, M. H., Jannat, M. K. A., Islam, M. S., Grossi, G., Bursic, S., & Aktaruzzaman, M. (2023). Real-time face mask position recognition system based on MobileNet model. *Smart health*, *28*, 100382.

Rajagukguk, N., Kencana, I. P. E. N., & Kusuma, I. G. L. W. (2024, May). Classification of Original and Fake Images Using Deep Learning-Resnet50. In *Proceedings of the First International Conference on Applied Mathematics, Statistics, and Computing (ICAMSAC 2023)*, 110, 51.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347-10357.

Yasser, B., Hani, J., El-Gayar, S., Amgad, O., Ahmed, N., Ebied, H. M., ... & Salah, M. (2023). Deepfake Detection Using EfficientNet and XceptionNet. In *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, 598-603.

Zhou, L., & Yu, W. (2022). Improved Convolutional Neural Image Recognition Algorithm based on LeNet-5. *Journal of Computer Networks and Communications*, *2022*(1), 1636203.

Bibliographic information of this paper for citing:

G, Akshatha; M, Kempanna; Ramanaiah, Preethi Kolluru; Doss, Bandi; S, Ramani & Rangasamy, Kirubakaran (2026). Data-Efficient Transformer Architectures for Image-Level Facial Forgery Detection: A Comparative Evaluation of ViT and DeiT. *Journal of Information Technology Management*, 18 (1), 17-33.
https://doi.org/10.22059/jitm.2026.106252