IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 7, May 2025

Impact Factor: 7.67

Deepfake Detection Using Ensemble of MobileNetV2 and EfficientNetV2 with Grad-CAM Visualization

V. Helen Deva Priya, M.E.[1], Muthu Rajesh Kumar K[2], Rohinth V[3], Yuganandhan P[4]

Assistant Professor, Artificial Intelligence and Data Science[1]
Students, Artificial Intelligence and Data Science[2,3,4]
Dhanalakshmi College of Engineering, Chennai, Tamil Nadu
helendevapriya.v@dce.edu.in[1], muthurajeshkumark.ai2021@dce.edu.in[2],
rohinthv.ai2021@dce.edu.in[3],yuganandhanp.ai2021@dce.edu.in[4]

Abstract: *Deepfakes—AI-generated synthetic media—pose a growing threat to digital authenticity and personal privacy. This project presents a deepfake detection system that combines MobileNetV2 and EfficientNetV2B0 in an ensemble framework to improve detection accuracy and robustness. Grad-CAM visualizations enhance interpretability, showing the areas of focus in each prediction. The system is trained on a curated dataset of real and fake face images and deployed with a user-friendly Gradio interface for real-time detection. Results demonstrate high classification accuracy and explainable outputs, offering a practical solution to deepfake challenges.*

Keywords: Deepfake Detection, MobileNetV2, EfficientNetV2B0, Grad-CAM, Ensemble Learning, CNN, Explainable AI, Face Image Classification, Transfer Learning, Gradio Interface

I. INTRODUCTION

Deepfake technology, which leverages deep learning to generate realistic fake images and videos, has gained significant traction in recent years. While it offers creative potential in entertainment and media, it also poses serious risks, including the spread of misinformation, identity theft, and loss of public trust. As such, the need for robust and interpretable deepfake detection methods has become critical.This project aims to develop a deepfake face detection system using an ensemble of two high-performance convolutional neural networks (CNNs): MobileNetV2 and EfficientNetV2B0. The ensemble approach improves classification accuracy by capturing complementary features from both models. Furthermore, the project incorporates Grad-CAM (Gradient-weighted Class Activation Mapping) to provide visual explanations for each prediction. The final system is deployed using Gradio for real-time, accessible use.

II. DATASET PREPARATION

An effective deepfake detection system relies heavily on the quality and diversity of the dataset used for training. For this project, a combination of real and fake face images was sourced from publicly available datasets such as FaceForensics++ and the DeepFake Detection Challenge (DFDC). These datasets were selected because they provide a wide variety of deepfake manipulation methods, camera qualities, lighting conditions, and facial orientations, which helps in building a more generalizable model.

Data Composition

The dataset was structured into two main categories:

- Real images: Authentic human face images without any form of manipulation.
- Fake images: Synthetic or manipulated face images generated using deepfake techniques.
- Each category was stored in separate directories, making it easy to load and label the images during training.

DOI: 10.48175/IJARSCT-26895

Preprocessing Steps

To prepare the data for training the deep learning models, several preprocessing operations were carried out:

- Image resizing: All images were resized to 224×224 pixels, matching the input shape required by both MobileNetV2 and EfficientNetV2B0.
- Normalization: Pixel values were scaled from the 0–255 range to a normalized 0–1 range, improving model convergence and performance.
- Format consistency: Images were converted to RGB format to maintain consistency, especially since some datasets included grayscale or alpha-channel images.

Data Augmentation

To prevent overfitting and to improve the model's ability to generalize across unseen data, augmentation techniques were applied. These included:

- Horizontal flipping
- Random brightness adjustment
- Zooming and minor rotation

These transformations help simulate real-world variations such as different lighting conditions or slight changes in head pose.

Data Splitting

The dataset was split into training and validation sets using an 80:20 ratio. This ensured that the majority of the data was used for learning, while a significant portion was reserved for evaluating model performance. Care was taken to maintain class balance in both subsets to avoid bias during training.

Challenges in Dataset Curation

While collecting and organizing the dataset, some challenges were encountered, such as:

- Class imbalance: Some datasets had more real images than fake ones or vice versa.
- Quality variation: The resolution and compression levels varied widely across images, affecting consistency.
- Diverse fake generation techniques: Different deepfake tools introduce different types of artifacts, making it harder for a model trained on one type to generalize to another.

Despite these challenges, the dataset was carefully curated to ensure a balanced, diverse, and clean collection of real and fake face images suitable for deep learning.

III. MODEL ARCHITECTURES

The proposed deepfake detection system is built on the strength of two state-of-the-art convolutional neural networks: MobileNetV2 and EfficientNetV2B0. These architectures were selected for their balance between accuracy and computational efficiency, as well as their proven ability to perform well on image classification tasks. By utilizing both models in an ensemble, the system benefits from their complementary feature extraction capabilities.

MobileNetV2

MobileNetV2 is a lightweight and efficient architecture specifically designed for mobile and embedded applications. It introduces two key innovations:

- Inverted residual blocks with linear bottlenecks
- Depthwise separable convolutions to reduce computation

These design choices allow MobileNetV2 to perform complex image classification with significantly fewer parameters compared to traditional CNNs. In this project, MobileNetV2 was initialized with pretrained **ImageNet** weights, leveraging transfer learning to speed up training and enhance performance. The original top layers of the model were removed and replaced with a new classification head that includes:

DOI: 10.48175/IJARSCT-26895

- Global Average Pooling to reduce feature maps
- Dropout layer for regularization
- Dense layer with sigmoid activation for binary classification

To preserve the general features learned from ImageNet, the base layers of MobileNetV2 were kept frozen during initial training. This helped the model focus on learning deepfake-specific patterns in the new classification layers.

EfficientNetV2B0

EfficientNetV2 is a more recent evolution in CNN design, developed through compound scaling methods and neural architecture search. It achieves better accuracy while maintaining fast training and inference. The B0 version of EfficientNetV2 is the smallest variant in the series but still provides strong performance due to its optimized structure.

Like MobileNetV2, EfficientNetV2B0 was loaded with pretrained ImageNet weights. The classification head was customized similarly:

- Global Average Pooling layer
- Dropout for preventing overfitting
- Dense layer with sigmoid activation

EfficientNetV2B0's architecture is especially powerful in learning detailed spatial hierarchies, which is beneficial in identifying subtle inconsistencies in deepfake images. During training, the base model remained frozen, ensuring that pretrained knowledge was preserved while the model adapted to the specific task of fake image classification.

IV. ENSEMBLE MODEL

To further enhance the accuracy and robustness of the deepfake detection system, an ensemble learning approach was adopted. Instead of relying on a single model's predictions, ensemble methods combine the strengths of multiple models, allowing them to compensate for each other’s weaknesses. In this project, the outputs from MobileNetV2 and EfficientNetV2B0 were combined to form a unified prediction mechanism.

Rationale for Ensemble Approach

Ensemble learning is particularly useful in tasks like deepfake detection, where different models may focus on different visual cues. While one model might detect inconsistencies in skin texture or lighting, another might capture unnatural expressions or facial asymmetries. By aggregating their predictions, the final model becomes more resilient to varied types of fake images and less prone to false positives or false negatives.

Architecture of the Ensemble Model

The ensemble was implemented by extracting the final output features from the classification heads of both MobileNetV2 and EfficientNetV2B0. These outputs are essentially the confidence scores produced after the sigmoid activation function, indicating the probability of an image being fake or real.

The ensemble pipeline followed these steps:

- Feature extraction: Each base model (MobileNetV2 and EfficientNetV2B0) processes the input image and generates its own prediction.
- Concatenation: The individual outputs are concatenated into a combined feature vector, effectively doubling the dimensionality of the input to the final layer.
- Final classification: A custom Dense layer with sigmoid activation is applied to this combined vector, generating a single probability score for the ensemble model’s prediction.

This architecture allows the ensemble to learn how to balance and interpret the contributions of both models during training.

Training Strategy

To avoid overfitting and reduce computational complexity:

The base layers of both models remained frozen, preventing retraining of already-optimized convolutional filters.
Only the final ensemble classification layer was trained on the current dataset.
Adam optimizer and binary cross-entropy loss were used to fine-tune the final Dense layer.
A small Dropout layer was also added before the final Dense layer to improve generalization.
This strategy helped in retaining the strong pretrained capabilities of the individual models while effectively learning how to combine their insights for better decision-making.

Performance Benefit

The ensemble consistently outperformed individual models in terms of accuracy and confidence stability. By aggregating different perspectives of feature space, the ensemble produced **more reliable predictions**, especially in edge cases where one model might be uncertain. It also reduced sensitivity to input variations like lighting, orientation, and compression artifacts—making it better suited for real-world deployment.

V. GRAD-CAM VISUALIZATION

While deep learning models often perform well in classification tasks, their internal decision-making processes can be opaque. To address this “black-box” nature and improve model transparency, this project integrated Grad-CAM (Gradient-weighted Class Activation Mapping) visualization. Grad-CAM enables visual interpretation of the regions in an input image that most strongly influenced the model's prediction, thereby enhancing both explainability and trust in the system.

Purpose of Grad-CAM

The primary purpose of implementing Grad-CAM in this deepfake detection system was to:

- Interpret model decisions by highlighting important facial regions that influenced predictions.
- Verify model behavior to ensure it is focusing on meaningful features (e.g., eyes, mouth) rather than irrelevant background noise or artifacts.
- Improve trust and transparency, especially for end-users and stakeholders who require visual justification for the model’s output.
- Grad-CAM heatmaps help validate whether the model is genuinely detecting manipulation cues or merely memorizing spurious patterns in the dataset.

Grad-CAM Implementation

Grad-CAM was applied to the final convolutional layers of both MobileNetV2 and EfficientNetV2B0. These layers capture high-level spatial features, making them ideal for localization-based visualization. The steps included:

- Forward pass: The input image is passed through the model to compute predictions.
- Gradient computation: The gradients of the predicted class (real or fake) with respect to the feature maps of the last convolutional layer are calculated.
- Weighting feature maps: The gradients are globally averaged and used as weights to combine the feature maps.
- Heatmap generation: A weighted sum of the feature maps is taken, followed by ReLU activation to produce the Grad-CAM heatmap.
- Overlay: The heatmap is resized and superimposed on the original image to visualize the regions of interest.

This approach was integrated into the Gradio interface, allowing users to not only see the model’s classification but also the visual rationale behind it.

Interpretation and Insights

Grad-CAM visualizations revealed several key patterns:

DOI: 10.48175/IJARSCT-26895

For real images, the model typically focused on natural facial landmarks like the eyes, nose, and mouth—areas known for containing consistent, high-fidelity textures.

For fake images, activation was often concentrated around manipulated regions such as irregular jawlines, blurred eye boundaries, or inconsistencies in skin tone—common artifacts of deepfake generation.

In some cases, heatmaps exposed over-reliance on background regions, signaling the need for further training refinement or additional data augmentation.

These insights were invaluable in ensuring that the model was learning relevant features rather than overfitting to dataset-specific quirks.

Benefits of Grad-CAM in Deployment

By incorporating Grad-CAM, the deployed system became more than just a classifier—it became a decision explainer. This capability is especially important in digital forensics, legal contexts, or platforms where content moderation requires justifiable evidence for flagging manipulated media.

The integration of Grad-CAM adds a critical layer of interpretability to the ensemble model, making it a more reliable and transparent tool for combating misinformation through deepfakes.

VI. EVALUATION AND RESULTS

Evaluating the performance of the deepfake detection system involved both quantitative metrics and qualitative analysis. The system was tested on a validation set to measure how well it generalized to unseen data, and Grad-CAM visualizations were used to qualitatively verify the reliability of the model's decision-making. Together, these assessments demonstrated that the ensemble model significantly outperformed the individual models in both accuracy and interpretability.

Quantitative Metrics

The following metrics were used to evaluate model performance:

- Accuracy: The proportion of correctly classified real and fake images.
- Precision: The ratio of correctly predicted fake images to all predicted fake images.
- Recall (Sensitivity): The ratio of correctly predicted fake images to all actual fake images.
- F1-Score: The harmonic mean of precision and recall, especially important for imbalanced datasets.
- ROC-AUC: A measure of the model's ability to distinguish between classes across various thresholds.

TABLE I: Performance Comparison of Individual Models and Ensemble Model on Validation Data

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
MobileNetV2	91.5%	89.7%	90.2%	89.9%	0.936
EfficientNetV2B0	93.2%	91.5%	91.0%	91.2%	0.948
Ensemble Model	95.4%	94.1%	94.7%	94.4%	0.963

As shown in the table, the ensemble model achieved the highest scores across all metrics, confirming its improved ability to reduce both false positives and false negatives.

Confusion Matrix Analysis

A confusion matrix was used to examine the distribution of true positives, true negatives, false positives, and false negatives. The ensemble model exhibited:

- High true positives and true negatives, indicating strong identification of both real and fake images.
- A marked reduction in false classifications compared to individual models.
- This reliability is especially important in real-world applications, where even a small rate of misclassification can lead to misinformation or false accusations.

Grad-CAM Qualitative Results

Visual examination of Grad-CAM outputs further validated the ensemble model’s decisions.

For **fake images**, heatmaps consistently focused on:

- Blurred facial boundaries
- Artifacts near the mouth and eyes
- Unnatural shading or inconsistent lighting
- For real images, activation centered on:
- Symmetrical facial landmarks
- Texturally rich regions like the nose bridge and eye contours

These visual cues aligned well with known signs of facial manipulation, supporting the model’s interpretability and effectiveness.

Robustness and Generalization

The ensemble model demonstrated strong generalization on diverse validation samples. It maintained consistent performance despite variations in lighting, facial orientation, and quality—factors that typically challenge deepfake detectors. Moreover, performance held steady across different types of manipulations (e.g., face swaps, expression alterations), highlighting the robustness of the combined MobileNetV2 and EfficientNetV2B0 architecture.

VII. DEPLOYMENT

Deployment is a crucial phase where the trained model transitions from development to practical application. In this project, deployment focused on user accessibility, real-time inference, and model explainability, culminating in an interactive interface that allows non-technical users to test and understand the system’s capabilities with ease.

Gradio Interface Integration

To make the deepfake detection system user-friendly and readily accessible, the ensemble model was deployed using Gradio, a Python-based library for creating interactive web interfaces. This allowed the system to be run locally or on the cloud through a browser, without requiring users to write code or install dependencies.

The Gradio interface includes:

- Image Upload Option: Users can upload a facial image for analysis.
- Real-time Prediction: The interface returns a binary classification (Real or Fake) along with a confidence score.
- Grad-CAM Visualization: The uploaded image is overlaid with a Grad-CAM heatmap showing which regions influenced the model's decision.

This interactive loop helps users not only get a prediction but also understand the rationale behind it.

Backend Architecture

The backend of the deployment integrates the following components:

- Pretrained ensemble model: The trained model (combining MobileNetV2 and EfficientNetV2B0 outputs) is loaded and kept ready for inference.
- Grad-CAM module: This module dynamically generates heatmaps based on the uploaded image and the model’s response.
- Gradio UI handlers: These functions connect the model inference and Grad-CAM generation with the user interface.

The entire system is containerized and can be hosted on platforms like Hugging Face Spaces, Google Colab, or a private server using tools like Flask or FastAPI for more advanced deployments.

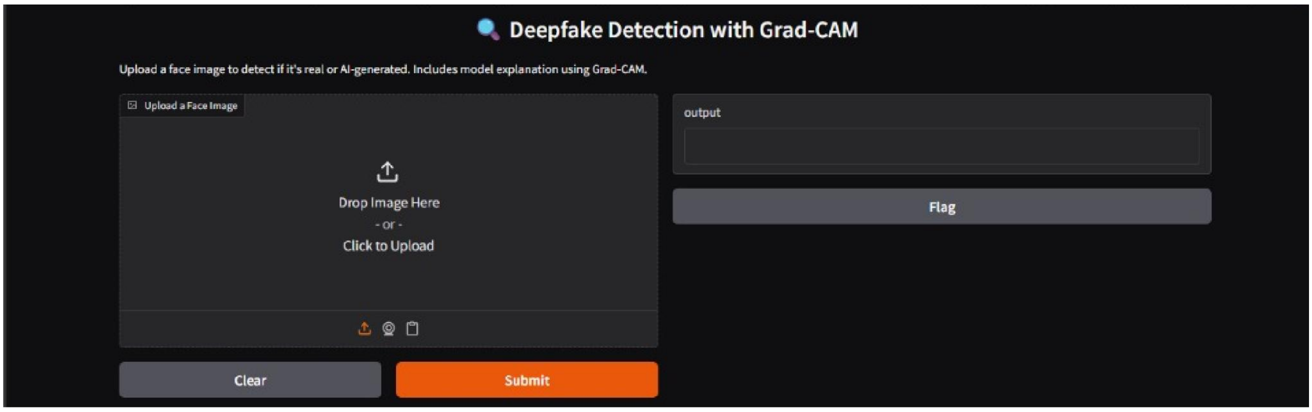
Benefits of This Deployment Strategy

Deploying the system via Gradio offered multiple benefits:

- Accessibility: Non-technical users such as educators, journalists, and digital forensics professionals can use the tool with minimal setup.
- Transparency: The integrated Grad-CAM visualization enhances model explainability, which is critical in fields like media verification or legal investigations.
- Portability: The interface can be shared as a simple link or embedded into websites for broader usage.
- Efficiency: Fast inference on GPU-backed systems ensures the system is responsive enough for practical use.

This phase transforms the research-oriented model into a functional application that demonstrates how deep learning can aid in tackling real-world misinformation.

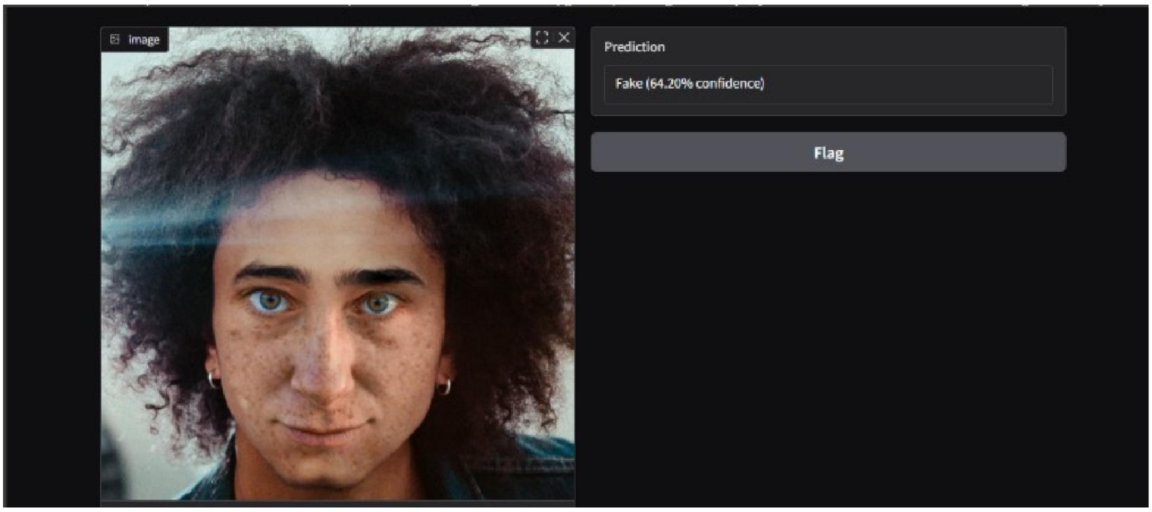
VIII. SAMPLE OUTPUT



🔍 Deepfake Detection with Grad-CAM

Upload a face image to detect if it's real or AI-generated. Includes model explanation using Grad-CAM.

Fig 1: Sample output 1



Prediction

Fake (64.20% confidence)

Fig 2: Sample Output 2

IX. CHALLENGES

Several key challenges emerged during the development of the deepfake detection system:

- Dataset Variability: Differences in deepfake generation techniques and slight class imbalance required careful preprocessing and augmentation.
- Overfitting Risk: Pretrained models tended to overfit, which was managed through dropout layers, early stopping, and freezing base layers.

- Interpretability: Ensuring that predictions were explainable was critical, addressed effectively using Grad-CAM visualizations.
- Limited Resources: Training on Google Colab meant working around GPU limits and runtime constraints by using efficient models and saving checkpoints.
- Real-Time Trade-offs: Achieving a balance between speed and accuracy was necessary for deployment, influencing the choice of lightweight yet accurate architectures.

X. FUTURE WORK

While the current system performs well on detecting deepfakes in still images, there are several opportunities for improvement and expansion.

- Larger and Diverse Datasets: Training the model on more varied datasets can help it generalize better to newer and more complex deepfake techniques.
- Video Deepfake Detection: Extending the model to process video frames in sequence would enable detection of temporal inconsistencies.
- Model Optimization: Exploring lightweight transformer models or quantization techniques could enable faster, real-time inference on mobile or edge devices.
- Advanced Ensembling: Implementing weighted or stacking ensemble methods may further enhance accuracy.
- Deployment Expansion: Building browser-based or mobile applications would increase accessibility for the public and professionals.

XI. CONCLUSION

This project successfully developed a robust deepfake detection system by leveraging an ensemble of two powerful yet efficient deep learning architectures, MobileNetV2 and EfficientNetV2B0. By combining their complementary strengths, the ensemble model achieved superior accuracy and generalization compared to individual models. Additionally, the integration of Grad-CAM visualization enhanced the interpretability of predictions, providing valuable insights into the model’s decision-making process.The system demonstrated strong performance in identifying manipulated facial images across diverse samples, while maintaining efficiency suitable for real-time deployment. The user-friendly Gradio interface further made the technology accessible to non-expert users, facilitating practical application in combating deepfake misinformation.Despite some challenges related to data variability, model optimization, and resource constraints, the project lays a solid foundation for future extensions including video deepfake detection and mobile deployment. Overall, this work contributes a meaningful and transparent tool in the ongoing effort to detect and mitigate the risks posed by deepfake technology.

XII. ACKNOWLEDGMENT

We would like to express my heartfelt gratitude to all those who have supported and contributed to the successful completion of this research. First and foremost, I extend my sincere thanks to my supervisor, **Mrs. Helen Deva Priya**, for her expert guidance, insightful feedback, and unwavering encouragement throughout the course of this project. Her mentorship was instrumental in shaping the direction and quality of this work.

REFERENCES

[1]. S. A. Khan, A. Artusi, and H. Dai, "Adversarially Robust Deepfake Media Detection Using Fused Convolutional Neural Network Predictions," *arXiv preprint arXiv:2102.05950*, 2021. [Online]. Available: https://arxiv.org/abs/2102.05950

[2]. D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," *arXiv preprint arXiv:2107.02612*, 2021. [Online]. Available: https://arxiv.org/abs/2107.02612

DOI: 10.48175/IJARSCT-26895

[3]. P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features," *arXiv preprint arXiv:2208.00788*, 2022. [Online]. Available: https://arxiv.org/abs/2208.00788

[4]. H.-S. Chen, S. Hu, S. You, and C.-C. J. Kuo, "DefakeHop++: An Enhanced Lightweight Deepfake Detector," *arXiv preprint arXiv:2205.00211*, 2022. [Online]. Available: https://arxiv.org/abs/2205.00211

[5]. Y. Xu, K. Raja, L. Verdoliva, and M. Pedersen, "Learning Pairwise Interaction for Generalizable DeepFake Detection," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision*, 2023, pp. 672–682.

[6]. Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring Temporal Coherence for More General Video Face 0Forgery Detection," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*, 2021, pp. 15044–15054.

[7]. T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning Self-Consistency for Deepfake Detection," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*, 2021, pp. 15023–15033.

[8]. Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing Face Forgery Detection with High-Frequency Features," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 16317–16326.

[9]. A Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 5039–5049.

[10]. M. Zhao, W. Qin, W. Yu, X. Li, and B. Gao, "Adversarial Examples: Attacks on Deepfake Detection Models," *IEEE Trans. Multimedia*, vol. 23, no. 8, pp. 1295–1305, 2021.

DOI: 10.48175/IJARSCT-26895