International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 1389-1397
https://doi.org/10.70917/ijcisim-2026-5136
Article

# RealVision – Deepfake Detection of Videos and Images : A Comparative Study of Custom CNN for Detection of Deep faked for Human Videos and Images

**Pushpa T. S.[1], S. Shilpa[2], B. R. Shreesha[3], Vignesh S.[4]**

[1–4] Department of Computer Applications, B.M.S. College of Engineering, Karnataka, India.
[1] Email: pushpa.mca@bmsce.ac.in
ORCID: 0000-0002-0328-8092
[2] Email: shilpa.mca@bmsce.ac.in
ORCID: 0009-0007-6745-7178
[3] Email: shreesha.mca23@bmsce.ac.in
[4] Email: vigneshs.mca24@bmsce.ac.in

**Abstract:** In the recent times rise in the content of the manipulated media even particularly deepfakes poses a serious threats to our digital integrity and privacy and the public trust. This study introduces a model called RealVision which is a lightweight and robust deepfake detection system capable of analyzing both images and video content of the humans. This designed to function efficiently on CPU-based systems. And the framework incorporates a custom-built Convolutional Neural Network (CNN) optimized for performance without relying on GPU acceleration keeping in mind the lower end devices . RealVision is trained using the Celeb DF V2 dataset and follows a multi-phase pipeline including dataset preprocessing, automated face extraction, model training and real-time inference. And also the evaluation metrics such as accuracy, precision, recall, and F1-score were used to assess performance across modalities. Our model's experimental results demonstrate competitive detection accuracy while maintaining low latency so making it suitable for the deployment in resource-constrained environments. RealVision takes into account to contribute to media authenticity verification tools thus in promoting digital safety in journalism, legal forensics, and public information platforms for betterment.
**Keywords:** RealVision, Deepfake, Custom CNN, Celeb DF V2,

## 1. Introduction

In recent years, the rapid growth of synthetic media, particularly deepfakes, has become a serious concern for digital security, online credibility, and public awareness. Deepfakes are images or videos that are manipulated using deep learning techniques to realistically imitate human faces, expressions, and voices. Because of their highly convincing nature, they are often difficult to detect through manual observation or traditional forensic techniques. The misuse of such manipulated content has raised global concerns, as it can lead to the spread of misinformation, character defamation, and even pose risks to national security.

Conventional detection mechanisms are increasingly proving inadequate to address the scale and sophistication of deepfake content circulating across social media platforms and communication networks. Manual analysis is time-consuming, prone to human bias, and impractical for large-scale or real-time applications. To address these challenges, recent advancements in artificial intelligence have enabled the development of automated systems for verifying media

authenticity. In particular, object detection and classification models can analyze visual patterns and identify subtle anomalies in facial consistency, texture, and motion that may not be easily noticeable to the human eye.

This study presents **RealVision**, a deep learning-based detection system designed to identify deepfake content in both static images and video frames. The proposed approach emphasizes efficient performance in low-resource environments, such as CPU-only systems, while still maintaining strong detection accuracy. By supporting both images and videos, the system provides a comprehensive solution that can be integrated into media verification pipelines, legal investigations, and journalistic practices.

Furthermore, this work aligns with the United Nations Sustainable Development Goals, particularly **SDG 9 (Industry, Innovation, and Infrastructure)** and **SDG 16 (Peace, Justice, and Strong Institutions)**. By contributing to the development of ethical AI tools for detecting manipulated media, the study promotes digital accountability and helps reduce the spread of harmful misinformation. Ensuring access to trustworthy information plays a key role in supporting peaceful societies and strengthening digital institutions across sectors such as law, governance, and public media. These broader social motivations form an important foundation for the development of the Real Vision system.

## 2. Literature Review

Deepfake detection has become an increasingly active area of research as synthetic media continues to evolve. Recent studies focus on developing lightweight and robust architectures, improving dataset quality, and incorporating explainability to enhance real-world detection performance. The following review summarizes six relevant studies and highlights how their findings relate to the development of the **RealVision** system.

**[1] Afchar et al.** proposed **MesoNet**, a mesoscopic convolutional neural network architecture designed to be both lightweight and efficient while still capturing subtle image patterns useful for detecting manipulated facial content in videos. The model focuses on identifying small visual inconsistencies that often appear in forged media. Their work also emphasizes robustness to video compression, which is important since most online content is heavily compressed. Experiments showed strong performance, reporting detection rates above 95% for some manipulation types, while achieving around **84% overall accuracy**. Their study demonstrates the balance between model complexity and computational efficiency, making it a valuable reference for lightweight detection systems.

**[2] Rössler et al.** introduced **FaceForensics++ (FF++)**, one of the most widely used benchmarks in deepfake research. This dataset contains a large collection of manipulated videos and approximately **1.8 million images**, generated using multiple face manipulation techniques such as DeepFakes, Face2Face, FaceSwap, and Neural Textures. Importantly, the dataset also includes various compression levels to simulate real-world social media conditions. Their work shows that models performing well on clean datasets may perform poorly when videos are compressed. This highlights the need for realistic evaluation conditions when developing deepfake detectors. Models evaluated on FF++ achieved around **81% accuracy** under these challenging conditions.

**[3]** Another recent study explored improvements to **XceptionNet-based architectures** for deepfake detection. The researchers introduced cross-attention mechanisms to better capture spatial and temporal relationships between video frames. They also applied **few-shot learning techniques** to improve the model's ability to detect previously unseen types of manipulations. Their results suggest that attention mechanisms can help the model focus on subtle inconsistencies across frames, while few-shot learning improves generalization to new manipulation techniques. However, these approaches require careful training to prevent overfitting. The proposed method achieved approximately **93% accuracy**.

**[4]** Some researchers have focused on deeper convolutional neural networks (D-CNNs) to extract more complex spatial features from manipulated media. These models often achieve very high accuracy on benchmark datasets, particularly when the input videos are high quality and uncompressed. In some studies, detection accuracy reached **98% on uncompressed datasets**. However, these models are typically more computationally expensive and may not perform as well when deployed in real-world scenarios with compressed or noisy data. This highlights the trade-off

between model performance and computational efficiency, which is an important consideration for systems designed to run on CPU-only environments.

**[5]** Dasgupta et al**.** propose integrating Squeeze-and-Excitation (SE) blocks into a lightweight CNNs to recalibrate channel-wise features. They take into account SE-enhanced small CNNs improve detection accuracy (reported ~94% on specific datasets) while keeping model size modest — a valuable design when deploying to low-resource/CPU environments. SE blocks are a lightweight way to boost discriminative power without large parameter increases. Achieved 94.1% accuracy

**[6 ]**The **ExDDV** dataset provides not only binary real or fake labels but also human-provided textual explanations and artifact location clicks to support explainable deepfake detection research. This resource highlights the importance of explanation and localization annotations for building trustworthy detectors and for validating explanation methods such as Grad-CAM or saliency maps. ExDDV demonstrates that explainability resources improve qualitative model evaluation and error analysis. They achieved 92.4% accuracy

## 3. Methodology

- For this study, we used the **Celeb-DF v2 dataset**, which contains a combination of real and manipulated face videos designed for deepfake detection research. The dataset provides realistic examples of synthetic media, making it suitable for training and evaluating deep learning models. Each video was preprocessed to extract meaningful frames that could be used as training samples. The main preprocessing steps are described below.
- Face Extraction:

Video frames were extracted at uniform intervals and processed using a pre-trained **MTCNN (Multi-task Cascaded Convolutional Network)** face detector. This step allowed us to isolate facial regions from each frame, since deepfake artifacts are most visible around the face area. Only detections with high confidence scores were retained to ensure that the dataset contained clear and reliable facial samples.

- **Dataset Split:**

After face extraction, the resulting images were divided into **training, validation, and testing sets** using a **70:15:15 ratio**. This split helps in properly evaluating model performance while avoiding overfitting. In total, **50,212 image frames** were used for training, **5,534 frames** for validation, and **6,068 frames** for testing.

- **Image Resizing & Normalization**:
    - All images were resized to **96×96** pixels to standardize input across the CNN model.
    - Pixel values were normalized between 0 and 1 to stabilize learning and improve convergence.
- **Data Augmentation**:
    - Applied techniques like **horizontal flip**, **zoom**, **rotation**, and **brightness adjustments** to enhance model robustness and simulate diverse real-world conditions.
    - Augmentation was carefully balanced to preserve critical facial features necessary for accurate deepfake detection.
- **Label Encoding**:
    - Real and fake samples were annotated using binary labels (0 for real, 1 for fake).
    - Class imbalance was mitigated by resampling strategies to ensure fair representation during training.

This carefully prepared and pre-processed dataset enabled our model to learn all possible patterns and artifacts often present in manipulated media forming the foundation of the RealVision detection pipeline.

## 4. Implementation

A modular architecture is followed in the implementation of Real Vision. First, we trained the keras model using the above stated dataset and then a web interface is developed using fast API and HTML and CSS. The web interface also provides frames analysed, graphical representation of result and time taken to analyse the video frames. Followings are the details for this

- **Frame Format:** RGB images, resized to **96×96 pixels**
- **Label Classes:** Binary (Real, Fake)
- **Augmentation:** Rotation (±10°), shifts (10%), horizontal flip
- **Preprocessing:** Pixel normalization (1./255), batch-wise loading to avoid memory overload

**Feature Extraction**

- Frames were sampled directly from videos using cv2.VideoCapture
- Each sampled frame was resized and normalized before prediction
- No explicit handcrafted features were used; learned features from CNN

**Model Architecture**

A lightweight, CNN-based architecture was designed for CPU efficiency:

- **Convolution Layers:** Multiple Conv2D blocks with ReLU activations and BatchNorm
- **Pooling:** Average Pooling and MaxPooling layers to reduce spatial size
- **Regularization:** L2 (1e-4) + Dropout (40%) to prevent overfitting
- **Final Layers:**
    - GlobalAveragePooling2D
    - Output: 2-unit Sigmoid layer for binary classification

**Training Setup**

- **Loss Function:** Binary Cross entropy
- **Optimizer:** Adam (learning rate = 3e-4)
- **Batch Size:** 16
- **Epochs:** 20 (with early stopping and best model checkpointing)
- **Callbacks:** TQDM progress, early stopping, confusion matrix plot

| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_layer (InputLayer) | (None, 96, 96, 3) | 0 |
| conv2d (Conv2D) | (None, 96, 96, 4) | 112 |
| batch_normalization (BatchNormalization) | (None, 96, 96, 4) | 16 |
| conv2d_1 (Conv2D) | (None, 96, 96, 8) | 296 |
| conv2d_2 (Conv2D) | (None, 96, 96, 8) | 584 |
| batch_normalization_1 (BatchNormalization) | (None, 96, 96, 8) | 32 |
| average_pooling2d (AveragePooling2D) | (None, 48, 48, 8) | 0 |
| conv2d_3 (Conv2D) | (None, 48, 48, 16) | 1,168 |
| conv2d_4 (Conv2D) | (None, 48, 48, 16) | 2,320 |
| conv2d_5 (Conv2D) | (None, 48, 48, 16) | 2,320 |
| batch_normalization_2 (BatchNormalization) | (None, 48, 48, 16) | 64 |
| average_pooling2d_1 (AveragePooling2D) | (None, 24, 24, 16) | 0 |
| conv2d_6 (Conv2D) | (None, 24, 24, 32) | 4,640 |
| conv2d_7 (Conv2D) | (None, 24, 24, 32) | 9,248 |
| conv2d_8 (Conv2D) | (None, 24, 24, 32) | 9,248 |
| batch_normalization_3 (BatchNormalization) | (None, 24, 24, 32) | 128 |
| average_pooling2d_2 (AveragePooling2D) | (None, 12, 12, 32) | 0 |
| conv2d_9 (Conv2D) | (None, 12, 12, 64) | 18,496 |
| batch_normalization_4 (BatchNormalization) | (None, 12, 12, 64) | 256 |
| max_pooling2d (MaxPooling2D) | (None, 6, 6, 64) | 0 |
| global_average_pooling2d (GlobalAveragePooling2D) | (None, 64) | 0 |
| dropout (Dropout) | (None, 64) | 0 |
| dense (Dense) | (None, 32) | 2,080 |
| leaky_re_lu (LeakyReLU) | (None, 32) | 0 |
| dropout_1 (Dropout) | (None, 32) | 0 |
| dense_1 (Dense) | (None, 2) | 66 |

**Figure 1.** Layers in RealVision

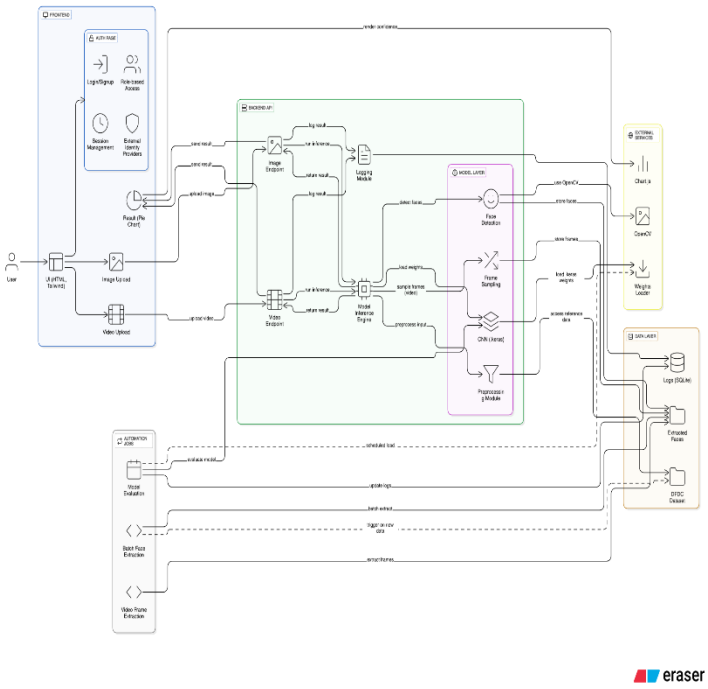

**Figure 2** System architecture for RealVision

## 5. Results

The performance of the proposed deepfake detection model Real Vision was evaluated using training and validation accuracy, loss trends, and confusion matrix metrics. The model showed a consistent upward trend in accuracy across epochs, reaching a peak validation accuracy of **93.46%** by the 6th epoch, as depicted in the training curves. Early stopping was employed at epoch 9 to prevent overfitting so with the model reverting to the best weights obtained at epoch 6

The **accuracy graph** indicated a steady increase in performance on both training and validation datasets, while the **loss curve** demonstrated a decline in both training and validation loss, further suggesting effective learning and convergence.

The confusion matrix highlights the model's strong capability to correctly classify real and fake video frames, demonstrating high precision in deepfake detection. Out of 5534 total validation samples:

- 3594 fake frames were correctly classified,
- 1578 real frames were accurately identified,
- 206 real frames were misclassified as fake,
- and 156 fake frames were misclassified as real.

The final evaluation on the hold-out test set yielded a test accuracy of **93.0%**, confirming the model's generalizability. And the training pipeline reported a final training accuracy of **96.01%** and a loss of **0.1241**, demonstrating robust model performance.

For **testing** the results were as follows

frames with high precision. Out of 6068 total **Testing** samples:

- 3618 fake frames were correctly classified,

- 2027 real frames were accurately identified,
- 293 real frames were misclassified as fake,
- and 130 fake frames were misclassified as real.

```
=== Classification Report ===
              precision    recall  f1-score   support

     fake(b)       0.93      0.97      0.94      3748
        real       0.94      0.87      0.91      2320

    accuracy                           0.93      6068
   macro avg       0.93      0.92      0.93      6068
weighted avg       0.93      0.93      0.93      6068
```

**Figure 3**. Evaluation metrics for RealVision

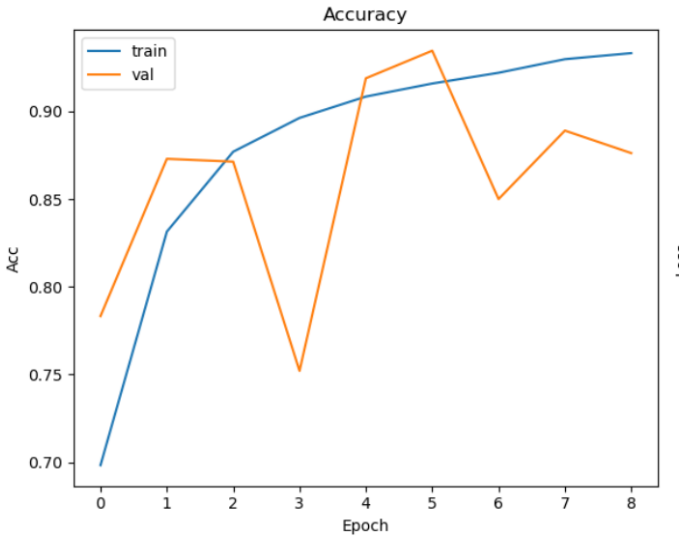


**Figure 4**. Training and validation graph

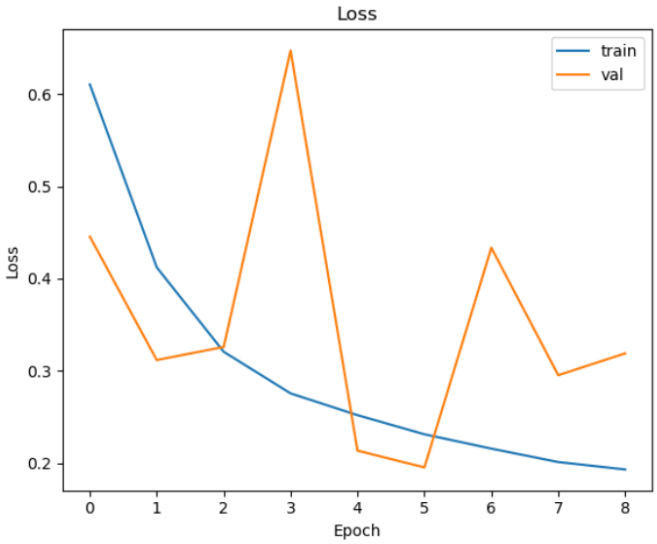


**Figure 5**. Data loss graph

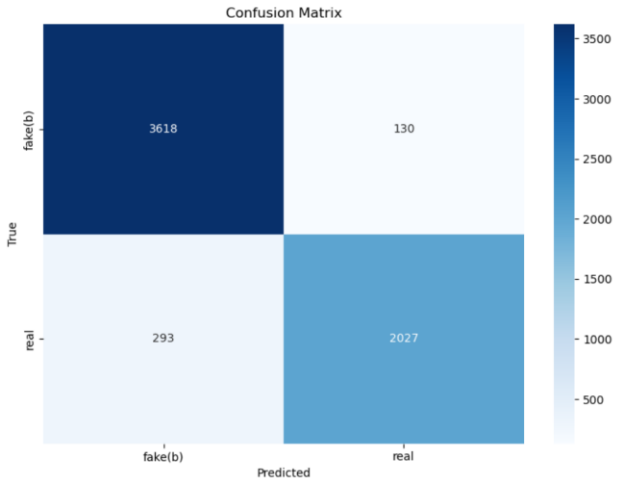


**Figure 6**. Confusion matrix

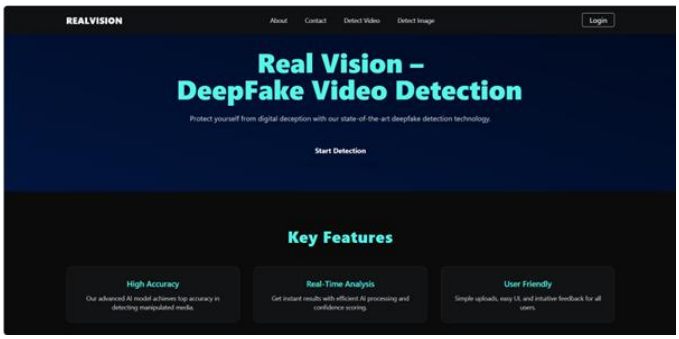


**Figure 7**. Home page

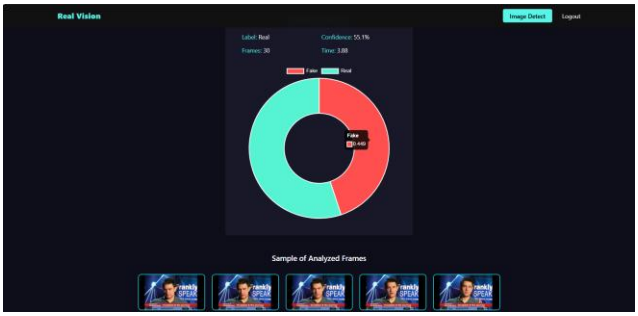


**Figure 8.** Result page


**Figure 9.** Admin dashboard

## 6. Conclusion

This study successfully demonstrates that the deep learning can be effectively used to detect deepfake content in both images and videos using a lightweight and optimized CNN. The RealVision model combines accuracy, speed and accessibility making it a practical tool for the deepfake detection even on the less resource available environments such as CPU only systems. The model was trained and evaluated on the Celeb DF V2 dataset and achieves reliable performance of 93% accuracy in testing data across various manipulated content while also offering the visual insights like confidence based pie charts and preview frames to improve the user experience.

Implementing a structured pipeline which ranges from dataset preprocessing and face extraction to model training, evaluation and web deployment the whole system was designed for real-world use and also focusing on automation, interpretability and responsiveness. Use of the custom CNN ensures a reduced computational load without sacrificing detection quality and accuracy. While the integration of image and video modules expands the RealVison's applicability across different content types.

From a broader perspective this work contributes meaningfully to the growing field of AI-based media forensics by offering the following implications:

- **Media Authenticity Verification:** RealVision Enables institutions, journalists, and platforms to verify the authenticity of the visual content.
- **Digital Literacy and Awareness:** RealVision raises the public awareness of the potential threats posed by the synthetic media and empowers the users with tools to identify manipulated content.
- **Cybersecurity and Law Enforcement:** Assists in identifying and mitigating malicious usage of the deepfakes in cybercrimes, identity theft and misinformation campaigns.
- **Research and Development:** Establishes a baseline for developing lightweight, scalable and real-time DeepFake detection models which is suitable for the deployment in resource constrained environments like mobile or embedded devices.
- **Ethical AI Deployment:** Encourages responsible and transparent use of AI by promoting detection mechanisms along with the generative models.

Visual interpretability tools such as confusion matrices and probability based visual outputs make the system more explainable and user-centric which provides transparency in predicting the outcomes. These features ensure that the model's decisions are not just accurate but also trustworthy.

**REFERENCES**


1. T. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a compact facial video forgery detection network," *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, China, Dec. 2018. doi: 10.1109/WIFS.2018.8630761
2. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," *arXiv preprint* arXiv:1901.08971, 2019. Available: https://arxiv.org/abs/1901.08971
3. S. Xie, S. Wang, and X. Yang, "Cross-Attention and Few-Shot Learning-Based Deepfake Detection Using XceptionNet," in *Proceedings of the International Conference on Advances in Electronics, Information and Communication (ICAEIC)*, 2025, pp. 168–174.
4. M. Sikandar, M. Khan, and S. Hussain, "Deepfake Detection Using Enhanced Deep Convolutional Neural Networks," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 9, pp. 1–8, 2020. doi: 10.14569/IJACSA.2020.0110901
5. A. Dasgupta, S. Dutta, and S. Bandyopadhyay, "Enhancing Deepfake Detection using SE-Block Attention Mechanism in Lightweight CNN Models," *arXiv preprint* arXiv:2303.00556, 2023. Available: https://arxiv.org/abs/2303.00556
6. P. Korshunov, A. Giuliani, and S. Marcel, "ExDDV: A New Dataset for Explainable Deepfake Detection and Verification," *arXiv preprint* arXiv:2207.06270, 2022. Available: https://arxiv.org/abs/2207.06270

**Dataset:** https://www.kaggle.com/datasets/reubensuju/celeb-df-v2