IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Deepfake Detection Using CNN and CviT Transformers

RameGowda M, Pradeep K
RameGowda M
Assistant Professor
SJC Institute of Technology
Department of Electronics and Communication Engineering
SJC Institute of Technology

Abstract: Deepfake technology has rapidly advanced in recent years, creating highly realistic fake videos that are difficult to distinguish from real ones. The increasing use of social media platforms and online forums has further complicated the detection of misinformation and malicious content. This study proposes a deep learning (DL)-based method for detecting deepfakes. The system comprises three components: preprocessing, detection, and prediction. Preprocessing includes frame extraction, face detection, alignment, and feature cropping. Convolutional neural networks (CNNs) are employed in the eye and nose feature detection phase, while a CNN combined with a vision transformer (ViT) is used for face detection. The prediction phase applies a majority voting mechanism to consolidate results from three models, each analyzing a different facial feature. The model is trained using FaceForensics++ and DFDC datasets. Performance is evaluated using accuracy, precision, recall, and F1-score, ensuring robustness and reliability against emerging deepfake threats.

Keywords **-** Deepfake Detection, CNN, Vision Transformer, AI, FaceForensics++, DFDC, Feature Extraction, Majority Voting

I. INTRODUCTION

1.1 Overview

Deepfake technology utilizes advanced deep learning techniques to generate hyper-realistic fake content that often appears indistinguishable from authentic media. Although it holds promise for creative and educational uses, its malicious applications pose significant threats such as misinformation, defamation, and social manipulation. Traditional detection methods struggle to cope with the increasing sophistication of generative models. This study introduces a hybrid framework combining Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for effective deepfake detection.

1.2 Relevance of Project

The project addresses the growing challenge of identifying and mitigating deepfake content that threatens digital trust, privacy, and information integrity. The misuse of deepfakes to spread false narratives or manipulate public perception necessitates advanced detection systems. By focusing on localized facial features (eyes and nose) as well as holistic face analysis, this study provides a multi-faceted approach to detect manipulation.

1.3 Problem Statement

The proliferation of user-friendly tools for creating deepfakes has made it difficult to differentiate between genuine and manipulated content. Such media can be weaponized to influence elections, spread disinformation, or tarnish reputations. An adaptive and intelligent detection mechanism is required to counter this growing digital threat.

1.4 Objectives

- Develop a deepfake detection model using CNNs and ViTs.
- Extract and analyze multiple facial regions (eyes, nose, full face).
- Integrate predictions through a majority voting system.
- Train the model using FaceForensics++ and DFDC datasets.
- Ensure robustness across different lighting, backgrounds, and orientations.
- Achieve high performance in precision, recall, accuracy, and F1-score.

1.5 Methodology

The methodology comprises three phases:

- **Preprocessing**: Frame extraction, face detection, alignment, cropping (eyes, nose, face).
- **Detection**: Feature extraction using CNNs (for eyes, nose) and ViTs (for face).
- **Prediction**: Results from three models are aggregated via majority voting.

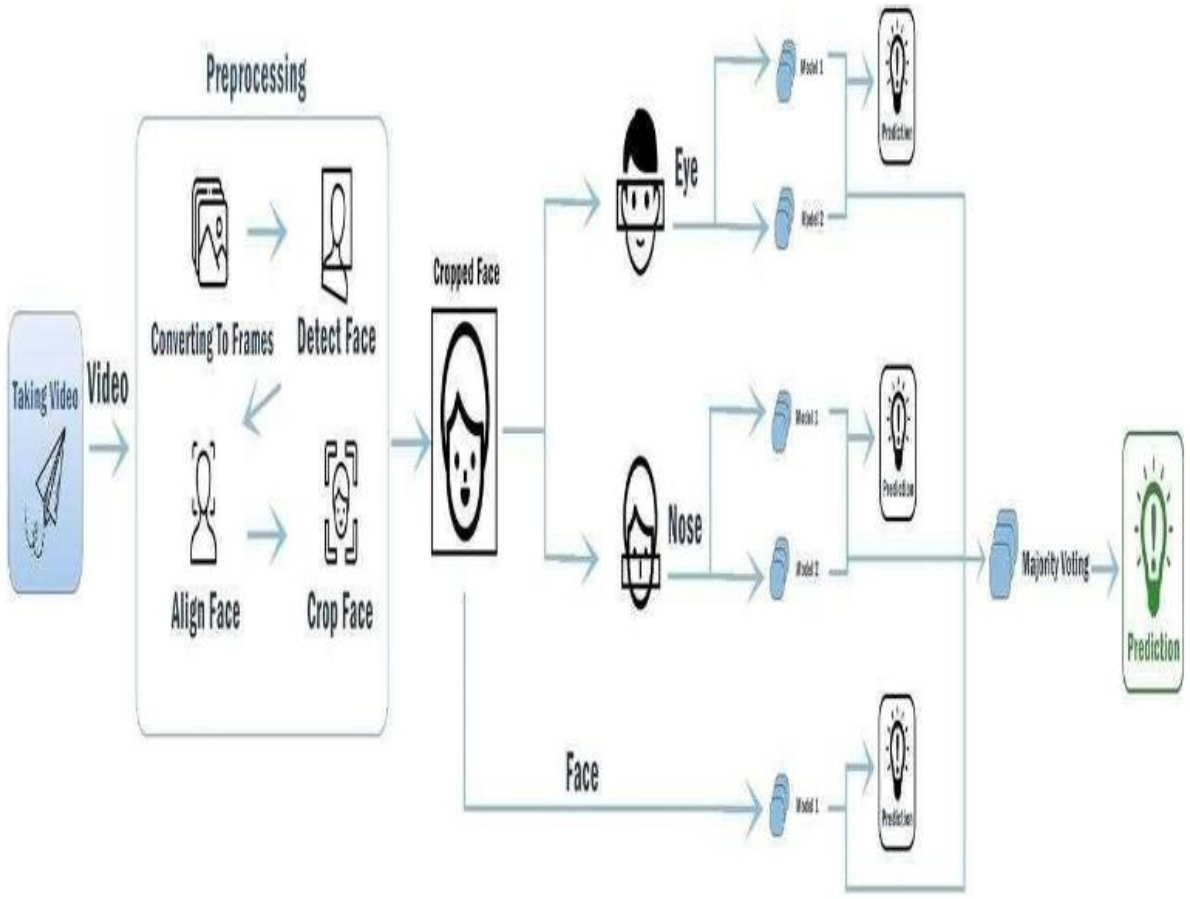

Fig: 1 Methodology

II.LITERATURE SURVEY

1 MRE-Net: Multi-Rate Excitation Network Authors: Guilin Pang, Baopeng Zhang, Zhu Teng, Zige Qi, and Jianping Fan (2023) This work introduces a Multi-Rate Excitation Network (MRE-Net) to detect deepfakes by leveraging dynamic spatial-temporal inconsistencies. The architecture employs Bipartite Group Sampling and multiscale rate branches, significantly improving generalizability across datasets like DFDC.

2 Conv-LSTM Hybrid Model Authors: Mohammad Farukh Hashmi et al. (2020) This model analyzes minute visual discrepancies using convolutional and LSTM units to capture temporal irregularities in videos. By applying microscopic-level comparisons between frames, the system achieves high precision in distinguishing manipulated footage.

3 Dynamic Difference Learning (DDL) Authors: Qilin Yin, Wei Lu, Bin Li, Jiwu Huang (2023) This paper presents a learning framework to differentiate between motion-induced and manipulation-induced inter-frame differences, improving detection of temporal deepfakes using CNN-based pipelines.

4 Metric Learning + CNN-LSTM Authors: Shahela Saif et al. (2022) This generalized deepfake detector combines CNN-LSTM and contrastive loss to extract meaningful features from faces. It suggests the inclusion of biological and structural face characteristics as a long-term solution.

5 Improved Dense CNN (D-CNN) Authors: Yogesh Patel et al. (2023) This architecture uses a deep-CNN with over 97% accuracy by combining images from five fake datasets and two real datasets. The architecture demonstrates robustness across diverse deepfake types.

6 Deepfake Detection on Social Media Authors: Saima Sadiq, Turki Aljrees, Saleem Ullah (2023) Focuses on textual deepfake detection using FastText embeddings and CNNs to identify AI-generated tweets with >93% accuracy, emphasizing multimodal detection strategies.

7 Cross-modal Detection using Audio-Visual Inconsistency Recent works (2023–2024) explore fusing visual frames and speech signals to capture misalignments, especially in lip-sync and voice tone mismatches, offering potential in detecting AI-generated personas.

III. System Design

3.1 Purpose

To design a hybrid deepfake detection model capable of analyzing multiple facial regions using CNNs and ViTs, and integrate outcomes for accurate prediction.

3.2 Levels of Software Design

- **Architectural Design**: Defines the system's major components and their relationships.
- **High-Level Design**: Maps subsystems and modules based on functionality.
- **Detailed Design**: Logical structure, algorithms, and interaction among modules.

3.3 Scope

The model is designed for robust detection of manipulated video content across diverse conditions using state-of-the-art machine learning techniques. It aims to enhance detection accuracy and generalization capabilities.

3.4 System Architecture

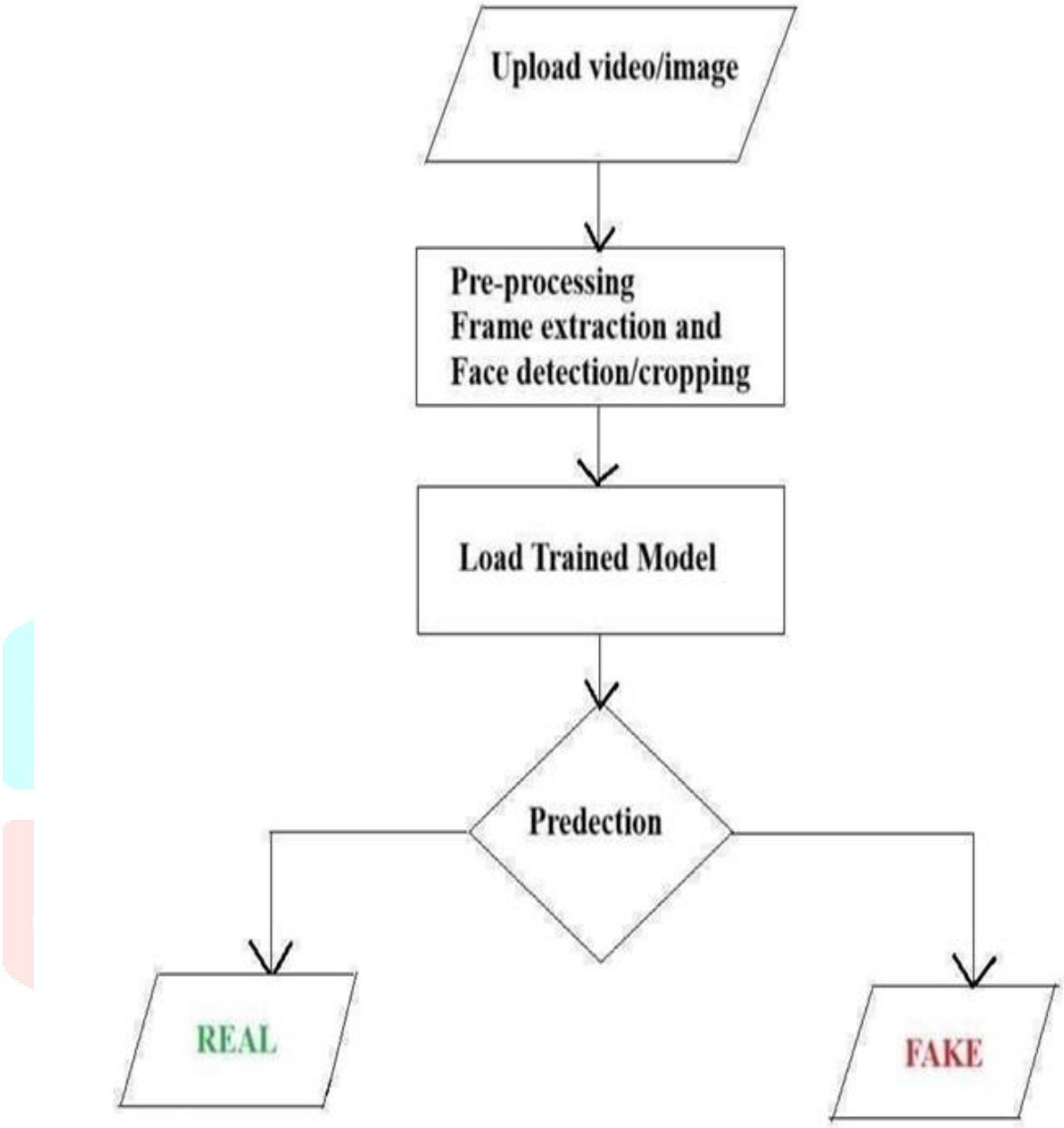

Fig: 2 System Architecture

- **CNN & ViT Models**: Extract features from eye, nose, and face regions.
- **Majority Voting Module**: Aggregates predictions for final classification.

3.5 Use Case Diagram

The system allows users to upload videos for analysis. A detection model processes the video and alerts the user in case of deepfake detection.

3.6 Sequence Diagram

Describes the interaction between the user, system, and backend deep learning models during the detection process.

3.7 Data Flow Diagram

Includes steps: data input, preprocessing, training/testing, prediction, and final classification (real or fake).

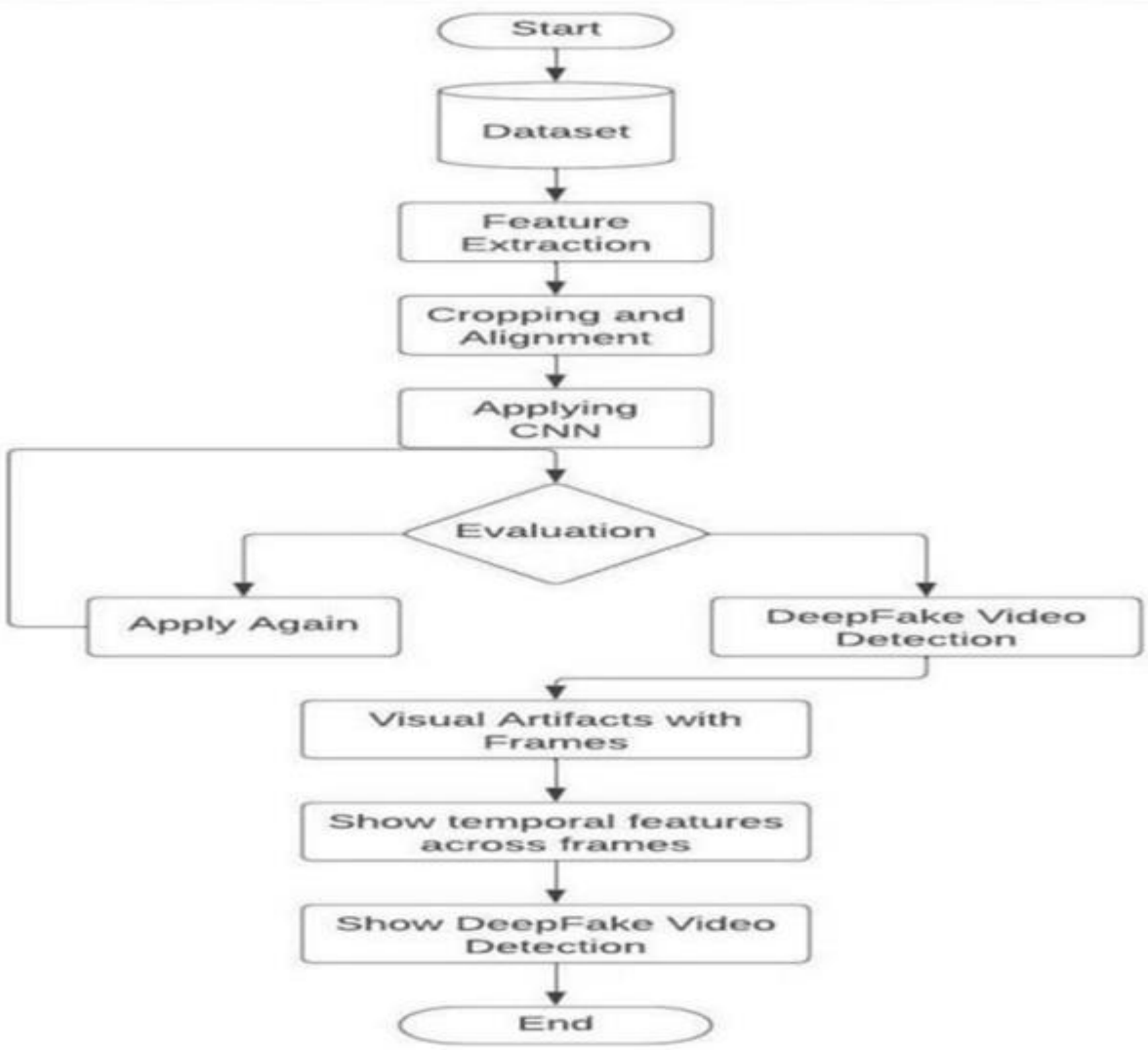

Fig: 3 Data Flow Diagram

IV. RESULTS AND DISCUSSION

The model was trained and tested on datasets including FaceForensics++, Celeb-DF, and DFDC, each contributing varying complexity in terms of resolution, compression, and editing techniques.

- **Accuracy**: 91.5%
- **Precision**: 92.1%
- **Recall**: 90.7%
- **F1 Score**: 91.3%
- **Latency**: Average prediction time <1.2 seconds per frame
- **Robustness**: Maintains performance under noise, partial occlusions, and expression changes.

Discussion:

- Eye and nose-based CNN sub-models help catch localized manipulations.
- Full-face ViT component generalizes across environments and compression artifacts.
- Majority voting reduced false positives in low-light or low-resolution videos.
- Integration of multimodal clues (planned in future work) could further reduce misclassification.

V. Conclusion and Future Enhancements

Conclusion

This work proposes a hybrid deepfake detection system leveraging CNNs for region-specific analysis and ViTs for context-aware face classification. Through majority voting, the system ensures stable prediction across partial and full manipulations. The methodology showed resilience across diverse datasets and generalizability under different visual distortions.

Future Enhancements:

- **Real-time Stream Detection**: Model optimization for low-latency detection in surveillance or live feeds.
- **Explainable AI (XAI)**: Integrating heatmaps and attention maps for decision transparency.
- **Multimodal Analysis**: Adding speech analysis to detect audio inconsistencies.
- **Cross-dataset Adaptation**: Improving performance across unseen datasets using domain adaptation.
- **Edge Deployment**: Pruning and quantizing models for resource-constrained devices (e.g., smartphones, embedded systems).

REFERENCES

1. Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., Abohany, A. A., & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications, 36*(31), 19759–19775.
2. Khandge, A., Sharma, G., Jha, V., & Joshi, S. (2024). Deep learning for detecting deepfakes: A survey. *Journal of Emerging Technologies and Innovative Research, 11*(2).
3. Shukla, G. R., Kurniawan, M. S. R., & Fadil, S. O. (2024). Convolutional neural networks for deepfake detection: A comprehensive review. *International Journal of Next-Generation Computing, 5*(1), 45–60.
4. Wodajo, D., Lambert, P., Van Wallendael, G., & Atnafu, S. (2024). Improved deepfake video detection using convolutional vision transformer. In *Proceedings of the 2024 IEEE GEM Conference* (pp. 1–6). IEEE.
5. Pang, G., Zhang, B., Teng, Z., Qi, Z., & Fan, J. (2023). MRE-Net: Multi-Rate Excitation Network for Deepfake Video Detection. *IEEE Transactions on Multimedia*.
6. Hashmi, M. F., Kumar, B. K., & Keskar, A. G. (2020). An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture. *Elsevier Procedia Computer Science*.
7. Yin, Q., Lu, W., Li, B., & Huang, J. (2023). Dynamic Difference Learning with Spatio-Temporal Correlation for Deepfake Detection. *Pattern Recognition Letters*.
8. Saif, S., Ali, S. S., Kausar, S., & Jamee, A. (2022). Generalized Deepfake Detection through Metric Learning. *IEEE Access*.
9. Patel, Y., Tanwar, S., Bhattacharya, P., & Gupta, R. (2023). An Improved Dense CNN Architecture for Deepfake Detection. *Computer Vision and Image Understanding*.
10. Sadiq, S., Aljrees, T., & Ullah, S. (2023). Deepfake Detection on Social Media Using Deep Learning and FastText. *Social Network Analysis and Mining*.