IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"DeepScan Pro - A Deepfake Detector"

[1]Miss. Shreya Jadhav, [2]Mr. Dhananjay Ambatwar, [3]Mr. Manasvi Mude, [4]Mr. Aryan Tapase, [5]Prof. Swapnil Powar

[1, 2, 3, 4]Student, [5]Assistant Professor

Department of CSE (Data Science),

D. Y. Patil College of Engineering & Technology, Kolhapur, India.

Abstract: "DeepScan Pro" is an advanced system designed to detect deepfake media, including images, videos, and audio, using state-of-the-art machine learning and deep learning algorithms. It analyzes subtle inconsistencies in media content by leveraging neural networks trained on real and deepfake datasets. The system adopts a multi-modal approach, allowing it to detect anomalies in images, videos, and audio by examining spatial features, motion patterns, and waveform inconsistencies. By incorporating advanced preprocessing techniques such as feature extraction, frame analysis, and spectral analysis, it ensures high accuracy and efficiency in deepfake detection. As the misuse of deepfake technology continues to pose significant risks in areas like cybersecurity, media integrity, and public trust, "DeepScan Pro" serves as a crucial defense mechanism. It has demonstrated high effectiveness in preliminary tests across diverse datasets, showcasing its ability to identify manipulated content with precision. The system's real-time processing capabilities make it highly adaptable for applications in digital forensics, media verification, and law enforcement. Designed for scalability, "DeepScan Pro" integrates an intuitive user-friendly interface, ensuring accessibility for users across various industries, including journalism, government agencies, and financial institutions. As deepfake technology evolves, the system continually updates its models with new data and emerging detection techniques to stay ahead of sophisticated manipulation methods. With its robust architecture, real-time analysis, and continual improvement, "DeepScan Pro" establishes itself as a powerful tool in the fight against deepfake deception, safeguarding the authenticity of digital media.

Keywords - Deepfake Detection, Deep Learning, Media integrity, Media Verification, Digital forensics, Public trust

I. INTRODUCTION

Deepfake technology has transformed digital media, enabling the creation of convincing fake content, including images, videos, and audio. While it has useful applications in entertainment, it also presents significant risks such as misinformation, fraud, and privacy violations. As deepfakes evolve, reliable detection methods are more crucial than ever. "DeepScan Pro" tackles this challenge by providing an advanced solution for detecting deepfake media. It uses state-of-the-art deep learning algorithms to accurately identify manipulated content across various media types, ensuring high reliability in detecting deepfakes and protecting digital integrity.

Need of the Work: The rapid evolution of deepfake technology has made it easier to manipulate media, creating serious concerns such as misinformation, fraud, and privacy violations. Traditional detection methods are no longer sufficient to combat these threats. An automated, AI-driven solution capable of real-time deepfake detection is necessary to preserve digital integrity and prevent the misuse of deepfake technology.

Existing Systems: Current deepfake detection tools, such as Deepware Scanner, Sensity AI, and Microsoft Video Authenticator, focus primarily on video analysis, with limited or no support for detecting manipulated images and audio. These systems often struggle with scalability, real-time processing, and accessibility, leaving gaps in effectively addressing the growing threats posed by deepfakes across different media formats.

Proposed System: "DeepScan Pro" presents a novel multi-modal deepfake detection framework that enhances current detection capabilities by integrating advanced deep learning techniques for improved accuracy and dependability. The main innovations of the proposed approach include:

- **Integrated Multi-Modal Detection –** Unlike conventional systems that focus on a single media type, "DeepScan Pro" performs simultaneous analysis of images, videos, and audio content. This holistic strategy ensures more thorough and reliable detection of manipulated media.
- **Use of Advanced Neural Architectures –** The system utilizes powerful deep learning models such as the Swin Transformer for visual analysis and CNN-based architectures for interpreting audio waveforms. These models are trained to recognize synthetic indicators like inconsistent facial expressions, abnormal lighting, and audio distortions.
- **Enhanced Preprocessing Pipeline –** To maximize detection precision, the system implements robust preprocessing methods including frame extraction, feature engineering, and spectral analysis. These techniques help isolate meaningful patterns and detect manipulation artifacts effectively.
- **Real-Time Analysis Capability –** The system is designed for low-latency performance, enabling quick deepfake verification. This makes it suitable for critical real-time applications in fields such as media verification, digital forensics, and online security.
- **Accessible User Interface –** A user-friendly interface built using Flet enables intuitive interaction. Users can easily upload media files and receive analysis results, making the system practical for both technical and non-technical audiences.

By leveraging these advancements, "DeepScan Pro" offers a scalable, accurate, and real-time deepfake detection solution, addressing the limitations of existing systems while enhancing digital security and trustworthiness.

II. LITERATURE SURVEY

Yan Wang, Qindong Sun, Dongzhu Rong, and Rong Geng [1] presented a specialized video deepfake detection technique aimed at handling compressed video content. Their approach incorporates adaptive notch filtering in the frequency domain, spatial noise reduction, and an attention-based fusion strategy to minimize compression-related distortions. However, its performance diminishes on uncompressed or lightly compressed media, affecting its ability to generalize across diverse datasets.

Jayashre K and Amsaprabhaa M [2] designed a composite framework named HODFF-DD by combining Inception ResNet with BiLSTM to detect both spatial and temporal irregularities in manipulated video clips. This architecture employs a spotted hyena algorithm to enhance feature extraction and time-based modeling. Although the system achieves strong outcomes, its focus on single-face video content restricts its applicability in more complex, multi-face scenarios.

Usamath Nechiyil, Nandakumar Paramparambath, Rahul TP, PR Aravind and Ranjith C [3] explored an audio-based deepfake detection model leveraging ResNet-34 with transfer learning techniques. Their work emphasizes the extraction of Mel-spectrogram features to detect synthetic speech, particularly in Automatic Speaker Verification (ASV) contexts. While the use of transfer learning enhances accuracy, its robustness under real-world conditions needs further exploration.

Rul Liu, Jinhua Zhang, and Guanglai Gao [4] proposed the MSCR-ADD framework to improve the accuracy of audio-based deepfake detection. This method transforms mono audio into binaural signals, capturing channel-specific and invariant features more effectively using multi-space representations. While this technique increases detection precision, it remains under-evaluated for detecting manipulated singing voices.

Hitesh Kumar Sharma, Manoj Kumar, and Preeti Sharma [5] developed a model that fuses GAN and CNN in an ensemble structure. The method utilizes generative replay to retain previously learned features, helping the system adapt to emerging deepfake variations. However, it lacks a comprehensive analysis of performance in complex manipulation environments.

Bo Wang, Fei Wang, Yushu Zhang, Fei Wei, Zengren Song, and Xiaohan Wu [6] introduced the Spatial-Frequency Fusion Branch (SFFB), which integrates spatial and frequency-based features using a knowledge transfer mechanism.This technique significantly improves detection on benchmark datasets such as FaceForensics++ and Celeb-DF. Despite its effectiveness, performance degradation under severe compression remains an unresolved issue.

Chen Li, Haoran Wang, and Xiang Gao [7] proposed a Transformer-based multi-modal detection model that simultaneously processes visual, audio, and video data. By learning interdependencies across modalities, the system enhances overall detection capability. While this approach surpasses single-modality techniques in performance, its complexity poses challenges for real-time execution, requiring further optimization.

III. METHODOLOGY

The methodology for **DeepScan Pro** focuses on detecting deepfake content in images, audio, and video. The approach combines multi-modal deep learning techniques to effectively detect subtle manipulations across different types of media. The key steps are:

1. **Data Collection**: A variety of datasets were carefully curated for training, testing, and validating the model. These datasets include real and AI-generated images, audio, and videos:
    - **Images**: 13,000 face images with a balanced split between real and AI-generated.
    - **Audio**: Real and synthetic audio clips in various Indian languages and accents, including datasets like ASVspoof and SceneFake.
    - **Videos**: High-resolution video datasets such as Celeb-DF V2 and a custom Hindi Audio-Video dataset with facial and vocal manipulations.
2. **Preprocessing**:
    - **Image Processing**: Facial regions were detected and cropped using DeepFace and face_recognition, followed by resizing and feature extraction using a **Swin Transformer**.
    - **Audio Processing**: Audio signals were converted into embeddings using **Wav2Vec2** and optionally analyzed with **MFCCs** or Mel-spectrograms.
    - **Video Processing**: Frames were extracted using **OpenCV**, resized, and passed through a **ResNeXt50** CNN for spatial feature extraction, followed by temporal pattern learning with an LSTM.
3. **Model Training**:
    - **Image Model**: The model uses **Swin Transformer** for feature extraction, followed by a custom neural network for classification.
    - **Audio Model**: Audio embeddings are passed through a **BiLSTM** model with attention to capture subtle speech manipulations.
    - **Video Model**: Frames are processed by **ResNeXt50** and fed into an LSTM to capture temporal inconsistencies in video content.
4. **Evaluation**: The system is evaluated using metrics like accuracy, ensuring reliable and unbiased deepfake detection across all media formats.
5. **User Interface**: A cross-platform user interface, built with **Flet**, enables real-time media uploads and deepfake detection with confidence scores displayed to users.

IV. SYSTEM DESIGN & IMPLEMENTATION

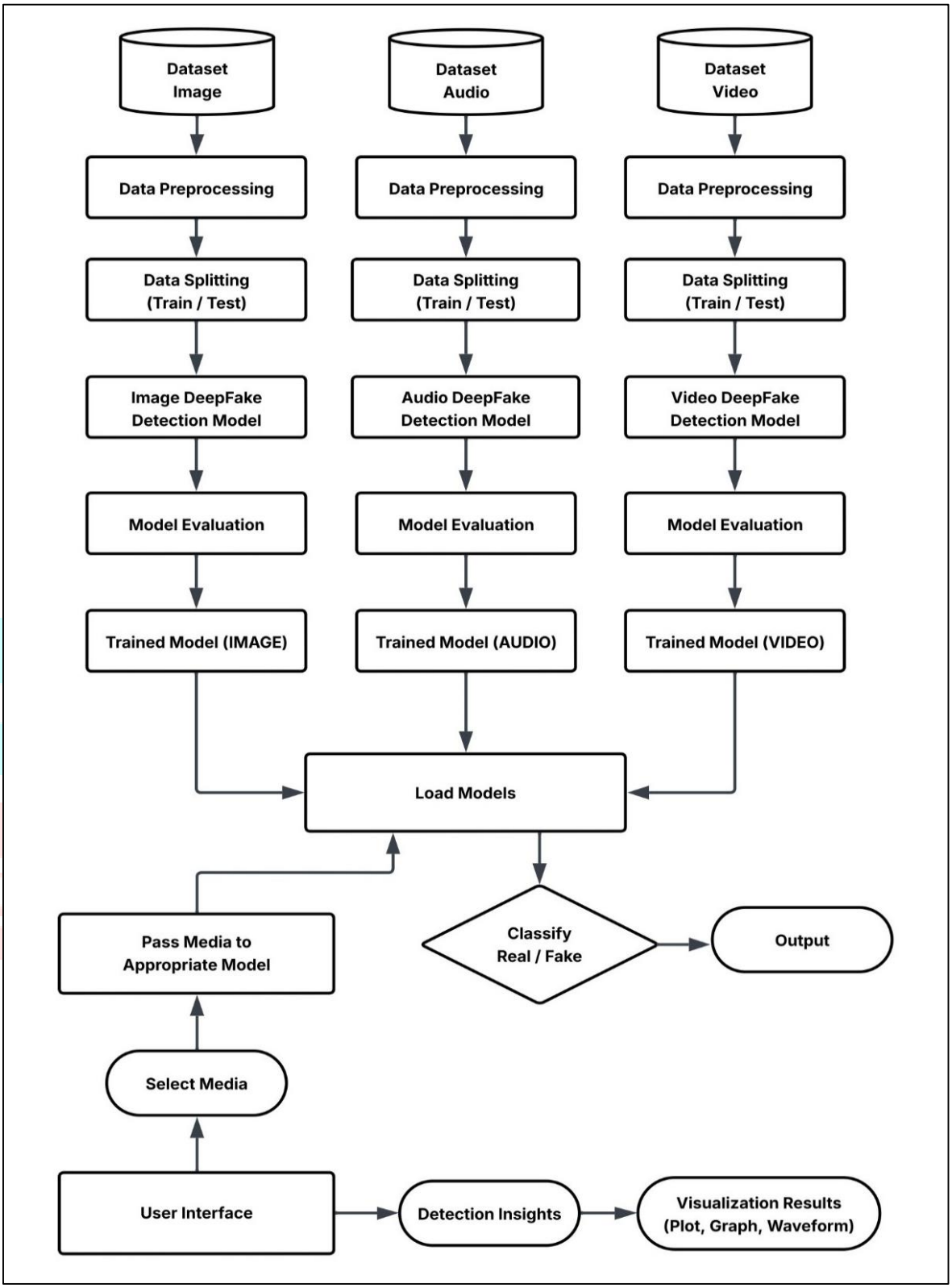


Fig a. System Architecture

Implementation Details: The DeepScan Pro system was developed using a modular and multi-modal architecture that combines deep learning techniques to detect deepfake content across multiple media types, including images, audio, and video. The system consists of several key components: data preprocessing, model training, prediction pipelines, and user interaction. The implementation is designed to handle real-time media analysis with high accuracy.

1. System Architecture:

- The system architecture is designed to handle diverse datasets (images, audio, video) collected from multiple domains, such as custom image datasets, face recognition datasets, CelebDF v2, Hindi audio-video datasets, and others. This ensures the system learns to identify the differences between genuine and tampered media. from various sources and formats.

- Data collection includes curated datasets that consist of both authentic and manipulated media samples. By integrating these datasets, the system is able to detect subtle patterns that differentiate real content from deepfakes in different media types.
- User Interface (UI): The system features a simple and intuitive UI built using Flet, which enables users to sign up, log in, and upload media files for analysis.

2. Data Preprocessing and Model Pipelines:

- Image Preprocessing: Facial regions are detected using face_recognition and DeepFace, then resized and normalized. Features are extracted using the Swin Transformer for robust spatial analysis.
- Audio Preprocessing: Audio signals are converted into Mel-spectrograms or Wav2Vec2 embeddings using Librosa. A BiLSTM with attention captures speech anomalies indicating deepfake manipulation.
- Video Preprocessing: Frames are extracted with OpenCV, resized, and processed. ResNeXt50 CNN captures spatial features, while LSTM models analyze temporal inconsistencies across frames.

3. Model Pipeline:

- Image Pipeline: After preprocessing, Swin Transformer extracts high-dimensional features, which are then classified using a custom neural network in PyTorch.
- Audio Pipeline: Audio features are first converted into embeddings via Wav2Vec2. These embeddings are then passed through a BiLSTM model with attention mechanisms to focus on subtle inconsistencies in speech patterns, such as pitch or timing changes.
- Video Pipeline: Each video frame is passed through ResNeXt50 to extract spatial features. The temporal aspect is captured using LSTM, which analyzes the sequence of frames and detects inconsistencies across the video.

4. User Interaction:

- Users can sign up, log in, and upload media through the interface.
- After selecting media, the system automatically detects its type (image, audio, or video) and processes it accordingly.
- The prediction is then made (Real or Fake), with the output displayed on the interface, showing the classification and confidence score.

5. Core Algorithm:

- **Input:** The input includes user credentials (for authentication) and the uploaded media file (image, audio, or video).
- **Output:** The system outputs a classification result, showing whether the media is Real or Fake based on the model's analysis.
- **User Flow:**
    1. Users are presented with options to sign up or log in.
    2. Upon logging in, users are directed to the Home Page where they can upload their media.
    3. The system identifies the media type and applies the corresponding preprocessing and feature extraction pipelines.
    4. The model generates a prediction (Real or Fake) and displays the result with a confidence score.
    5. Users can log out when finished.

6. Technological Stack:

- Python is used as the main programming language, leveraging powerful libraries such as PyTorch, TensorFlow, OpenCV, and Librosa for model development, feature extraction, and media preprocessing.
- **Firebase** is employed for user authentication and database management, providing secure and efficient user registration and data storage.
- **Flet** is used for the user interface, offering a cross-platform solution for seamless interaction with the system.

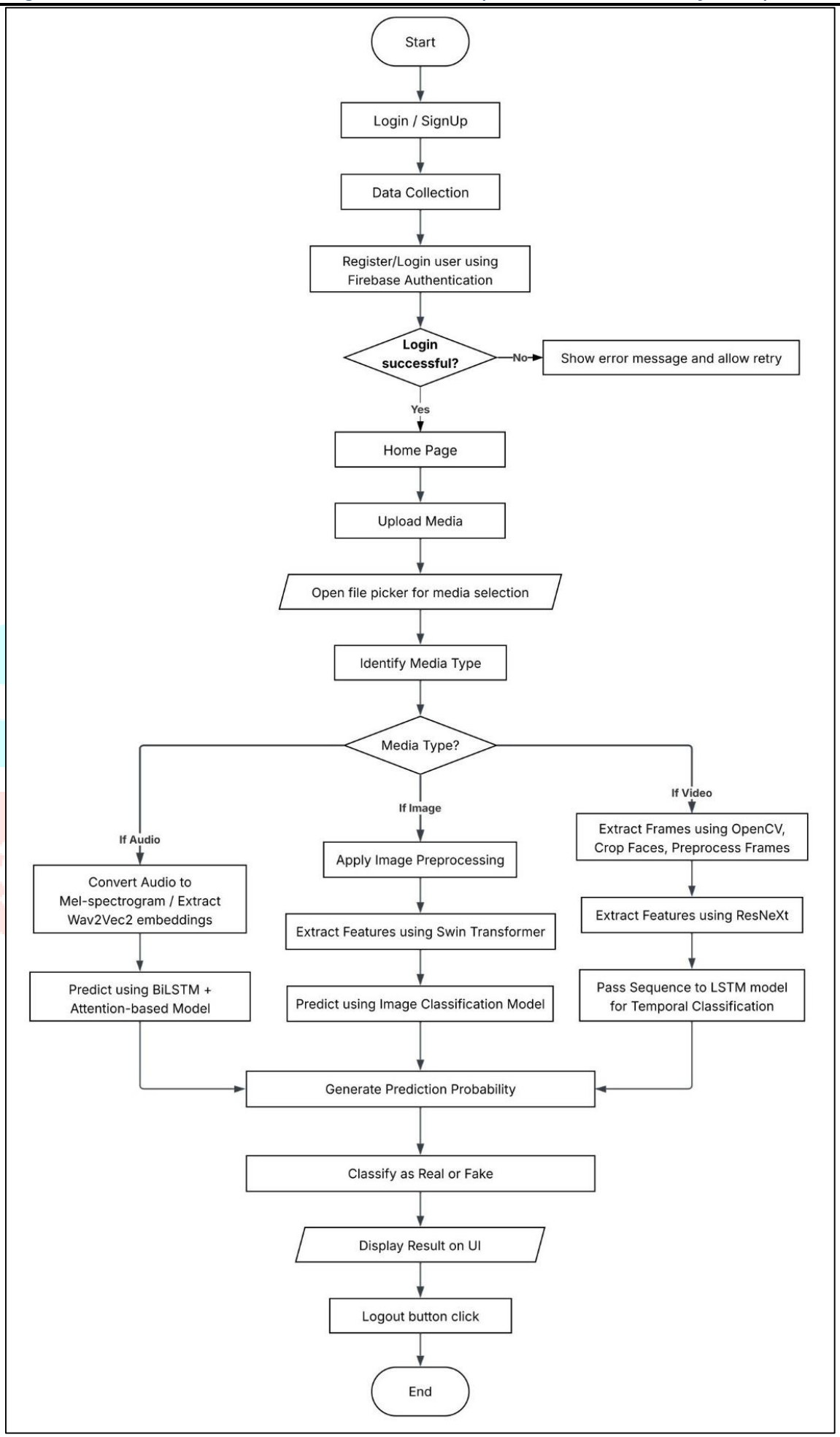


Fig b. Implementation Diagram

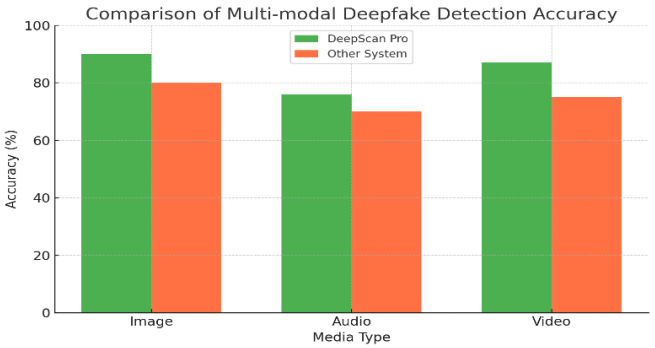
V. Result Analysis

The proposed multi-modal deepfake detection framework displayed strong and consistent performance across the image, audio, and video domains. Each module was independently trained and evaluated using datasets specific to its input type to ensure fair and accurate testing conditions. The system's effectiveness was evaluated using standard metrics such as accuracy.

Evaluation outcomes demonstrated that both the individual models and the integrated system achieved dependable results in detecting manipulated media. Specifically, in image classification, the Swin Transformer substantially outperformed traditional models like ResNet18 and EfficientNetB4, achieving 90% accuracy. While ResNet18 reached 67% and EfficientNetB4 scored 75%, the Transformer model showed superior consistency, justifying its selection for this component.

For audio deepfake identification, the Wav2Vec2 model significantly outclassed a baseline convolutional network. The CNN reached an accuracy of 64%, whereas Wav2Vec2 improved detection to 76% by leveraging pretrained feature representations that enhanced its adaptability to diverse and noisy audio inputs. Its capability to generalize across languages further solidified its selection.

In the video domain, a hybrid architecture that combined ResNeXt and LSTM networks attained an accuracy of 87%. This model captured both spatial and temporal elements, making it effective for detecting video-based manipulation. When compared with existing tools like Deepware Scanner, the proposed system—DeepScan Pro—exhibited better performance across all modalities, highlighting its strength in terms of precision, flexibility, and suitability for near-real-time detection.



VI. Conclusion

DeepScan Pro has proven to be a robust, multi-modal deepfake detection system capable of analyzing and identifying manipulated media across images, audio, and video. By utilizing advanced deep learning models like Swin Transformer, Wav2Vec2 with BiLSTM, and ResNeXt50 with LSTM, the system effectively captures spatial, spectral, and temporal anomalies that are characteristic of deepfakes. The integration of a user-friendly interface and real-time analysis features ensures that the system is practical, efficient, and suitable for real life applications, particularly in areas like journalism, digital forensics, and law enforcement.

In comparison with existing systems, DeepScan Pro addresses key limitations such as limited modality coverage, high hardware dependencies, and restricted access. With its diverse dataset strategy—including region-specific content—it delivers higher accuracy and contextual relevance. The system's modular architecture and explainable outputs further enhance trust and transparency, making it a reliable solution for detecting deepfake threats in evolving digital landscapes.

VII. References

[1] Wang, Y., Sun, Q., Rong, D., & Geng, R. (2024). Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts. Computer Vision and Image Understanding, 104(2), 72-81. ScienceDirect. https://doi.org/10.1016/j.cviu.2023.103852

[2] Jayashre, K., & Amsaprabhaa, M. (2024). Safeguarding media integrity: A hybrid optimized deep feature fusion-based deepfake detection in videos. Computers & Security, 103(4), 860-868. ScienceDirect. https://doi.org/10.1016/j.cose.2024.103520

[3] Rahul, T.P., Aravind, P.R., Ranjith, C., Nechiyil, U., & Paramparambath, N. (2024). Audio spoofing verification using deep convolutional neural networks by transfer learning. arXiv preprint arXiv:2008.03464, 28(2), 34-42. ScienceDirect. https://doi.org/10.48550/arXiv.2008.03464

[4] Liu, R., Zhang, J., & Gao, G. (2024). Multi-space channel representation learning for mono-to-binaural conversion-based audio deepfake detection. Information Fusion, 102(2), 257-264. ScienceDirect. https://doi.org/10.1016/j.inffus.2024.03.009

[5] Sharma, P., Kumar, M., & Sharma, H.K. (2024). GAN-CNN Ensemble: A robust deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique. Procedia Computer Science, 204(5), 90-99. ScienceDirect. https://doi.org/10.1016/j.procs.2024.02.016

[6] Wang, B., Wu, X., Wang, F., Zhang, Y., Wei, F., & Song, Z. (2024). Spatial-frequency feature fusion-based deepfake detection through knowledge distillation. Journal of Information Fusion, 65(1), 257-268. ScienceDirect. https://doi.org/10.1016/j.inffus.2024.01.005

[7] Li, C., Wang, H., & Gao, X. (2023). Multi-modal deepfake detection using Transformer-based cross-modal fusion. Journal of Multimedia Forensics and Security, 58(2), 189–202. Elsevier. https://doi.org/10.1016/j.jmfs.2023.189