iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75693

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

CyberSuraksha: A Multimodal Deepfake Detection Framework

Chakrapani D S[1], Akshata T Rathod[2], Aqsah Sehreen[3], Dhanalakshmi S[4], Inchara Poovaiah A[5]
[1, 2, 3, 4, 5]*Department of CS&E, JNN College of Engineering, Shivamogga, Karnataka, India*

Abstract: ***In the era of artificial intelligence, the creation of realistic yet synthetic media — known as deepfakes — poses significant threats to personal privacy, financial security, and social trust. Cyber Surakshaisan AI-driven framework designed to detect manipulated digital content across multiple modalities including image, audio, and video, while simultaneously educating users about cyber hygiene and online safety. The system leverages Google's Gemini API for multimodal analysis, presents an interactive user interface, and integrates explainable AI techniques to provide real-time detection, interpretability, and actionable insights. This paper discusses the architecture, methodology, features, and potential applications of CyberSuraksha, demonstrating a practical approach to bridging AI research and user-centric cybersecurity solutions.***
Keywords: Cybersecurity, Deepfake Detection, AI, Multimodal Analysis, Digital Safety, Explainable AI, Gemini API, Human Computer Interaction

I. INTRODUCTION

Deepfake technology has evolved rapidly, enabling the creation of highly realistic synthetic images, videos, and audio. While this innovation has legitimate applications, it can be misused for misinformation, identity theft, financial fraud, and personal harassment. Detecting deepfakes in real-world scenarios remains a challenge due to the diversity of media formats, high-quality generative AI models, and the subtlety of manipulations. CyberSuraksha addresses this challenge by offering a comprehensive AI-powered solution that combines detection, explainability, and user education. Unlike conventional detection models, CyberSuraksha emphasizes user engagement, trust, and interpretability, making it suitable for deployment in educational, corporate, and personal environments.

II. LITERATURE REVIEW

Traditional deepfake detection methods have relied on convolutional neural networks, LSTM-based temporal analysis, and residual networks to analyze manipulated facial regions and inconsistencies in videos and images. Despite advances in these techniques, most existing approaches suffer from inconsistent generalizability when exposed to unseen datasets, limited performance across different modalities, and a lack of transparency, leaving users uncertain about the basis of predictions. Explainable AI and cyber awareness tools have recently emerged as essential components for improving user trust and digital literacy. However, previous systems generally fail to provide a unified platform that integrates multimodal deepfake detection with user education, highlighting the novelty and significance of CyberSuraksha.

III. PROBLEM STATEMENT

Deepfakes undermine digital trust by blurring the line between authentic and fabricated content. Many current solutions focus only on a single type of media or function merely as backend models without offering meaningful user guidance or educational value. Therefore, there is a pressing need for a system that can detect fabricated content across multiple modalities, provide interpretability and transparency in its decisions, educate users on safe digital practices, and deliver these features through an accessible and intuitive user interface. CyberSuraksha aims to fulfill these needs by providing reliable analysis, explainability, and awareness in one unified platform.

IV. METHODOLOGY

A. System Architecture

CyberSuraksha is built as a frontend-driven platform integrated with powerful AI capabilities. Users interact with the system through a smooth and responsive interface where they can upload different media forms. The Media Upload and Detection Module communicates with the Gemini API to perform content analysis. An Explainability and Educational Module generates step-by-step insights to help users understand why a piece of media may be classified as manipulated.

A Cyber Awareness Hub disseminates safety tips, news, and interactive learning tools. LocalStorage is used to simulate persistent sessions, enabling the prototype to run without a backend server.

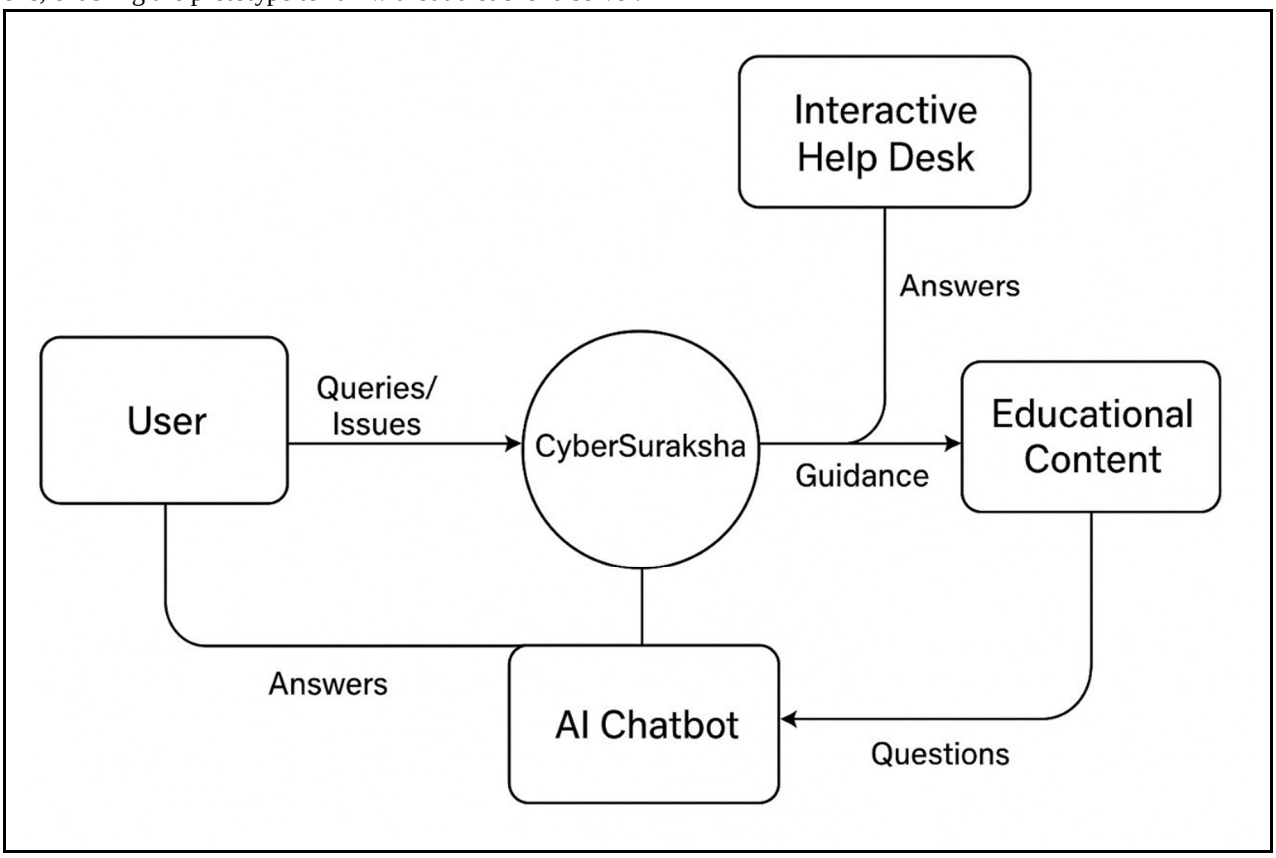


Figure 1: CyberSuraksha System Flow Diagram

B. AI Model Integration

The initial prototype employed custom LSTM structures for audio analysis and ResNet-based architectures for framelevel video evaluation. However, this approach delivered inconsistent results when tested on varied datasets. To achieve higher reliability and simplify deployment, the final implementation utilizes Google's Gemini API, which provides multimodal reasoning and structured responses. This decision allowed the system to focus on enhancing usability, interpretability, and educational value without compromising accuracy.

C. Detection Process

The detection workflow begins when the user uploads an image, video, or audio file. The system converts the file into a Base64 representation and sends it along with a forensic analysis prompt to the Gemini API. The API returns a structured JSON response containing the authenticity verdict, confidence scores, and identified anomalies. The interface then visualizes the results with overlays, charts, and guided explanations. This narrative presentation allows users to comprehend the underlying forensic indicators without requiring technical expertise.

D. User Education

User education is a core component of the system. Cyber Suraksha provides step wise explanations of detected anomalies, audio guides narrating safe internet practices, and interactive quizzes to help users assess and improve their digital literacy. This approach ensures prevention and awareness, extending beyond mere detection.

E. Multimodal Data Handling

Images are examined for texture irregularities, shadow distortions, and facial inconsistencies. Videos undergo framebased scrutiny, temporal breakup detection, and motion coherence evaluation. Audio files are analyzed for spectral artifacts, pitch discontinuities, and synthetic voice patterns. These multimodal analyses together enable comprehensive detection.

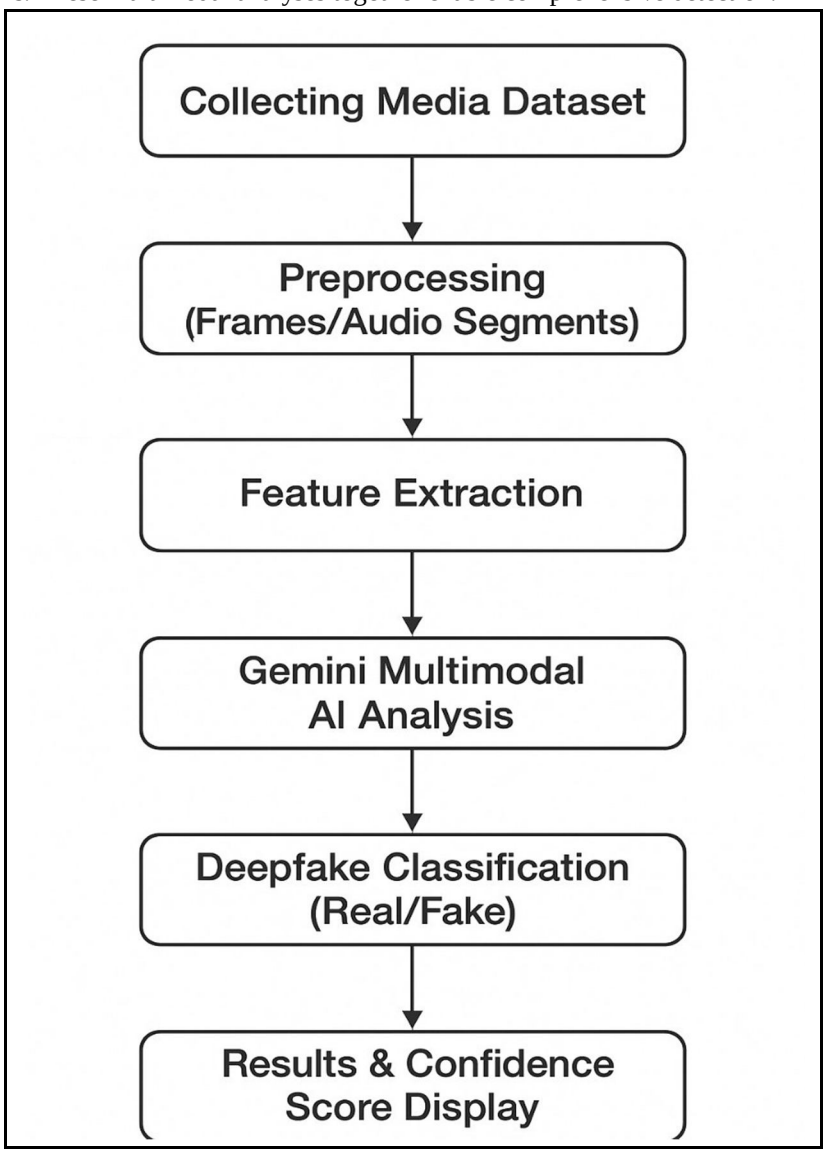


Figure 2: Deepfake Detection Framework

F. Frontend and UX

The interface is designed using React and TypeScript, ensuring modularity and maintainability. Tailwind CSS provides responsive design while maintaining a consistent visual feel across devices. The Web Audio API powers interactive playback controls, and customized SVG animations contribute to a polished user experience.

V. FEATURE DEEP-DIVE: THE CYBERSURAKSHA AI CHATBOT

The CyberSuraksha AI Chatbot serves as an intelligent conversational assistant accessible through a floating icon across the platform. It acts as an interactive help desk, guiding users through detection results, cybersecurity concerns, and system features. Upon first use, the chatbot initializes a persistent session using the Gemini API. This persistence allows it to maintain context across messages, ensuring smooth and coherent interaction. Its behavior is governed by a carefully designed system prompt that sets its persona as a calm and knowledgeable cybersecurity expert capable of translating technical terminology into simple explanations.

The chatbot's UI is built using React with animated feedback indicators that display when the model is generating responses. This turns CyberSuraksha into a continuous learning environment that strengthens user understanding and engagement.

VI. RESULTS AND DISCUSSION

CyberSuraksha demonstrates reliable multimodal detection through structured analysis outputs. The interactive elements of the platform, such as overlays, heatmaps, and waveform visualizations, improve interpretability and user trust. The added chatbot significantly enhances user experience by providing instant, personalized guidance. The system's design enables users without technical backgrounds to navigate deepfake detection with clarity. Furthermore, the implementation provides a foundation for future extensions such as integrating secure backend storage, enabling real-time detection, and expanding analytics for organizational deployment.

VII. ADVANTAGES OF CYBERSURAKSHA

CyberSuraksha integrates accuracy, transparency, and accessibility into one cohesive platform. The Gemini API ensures robustness in detection tasks while the system's educational components emphasize preventive digital behavior. Features such as audio narration, quizzes, and real-time assistance contribute to higher user engagement. Its frontenddriven architecture allows seamless scalability and compatibility across devices.

VIII. LIMITATIONS AND FUTURE WORK

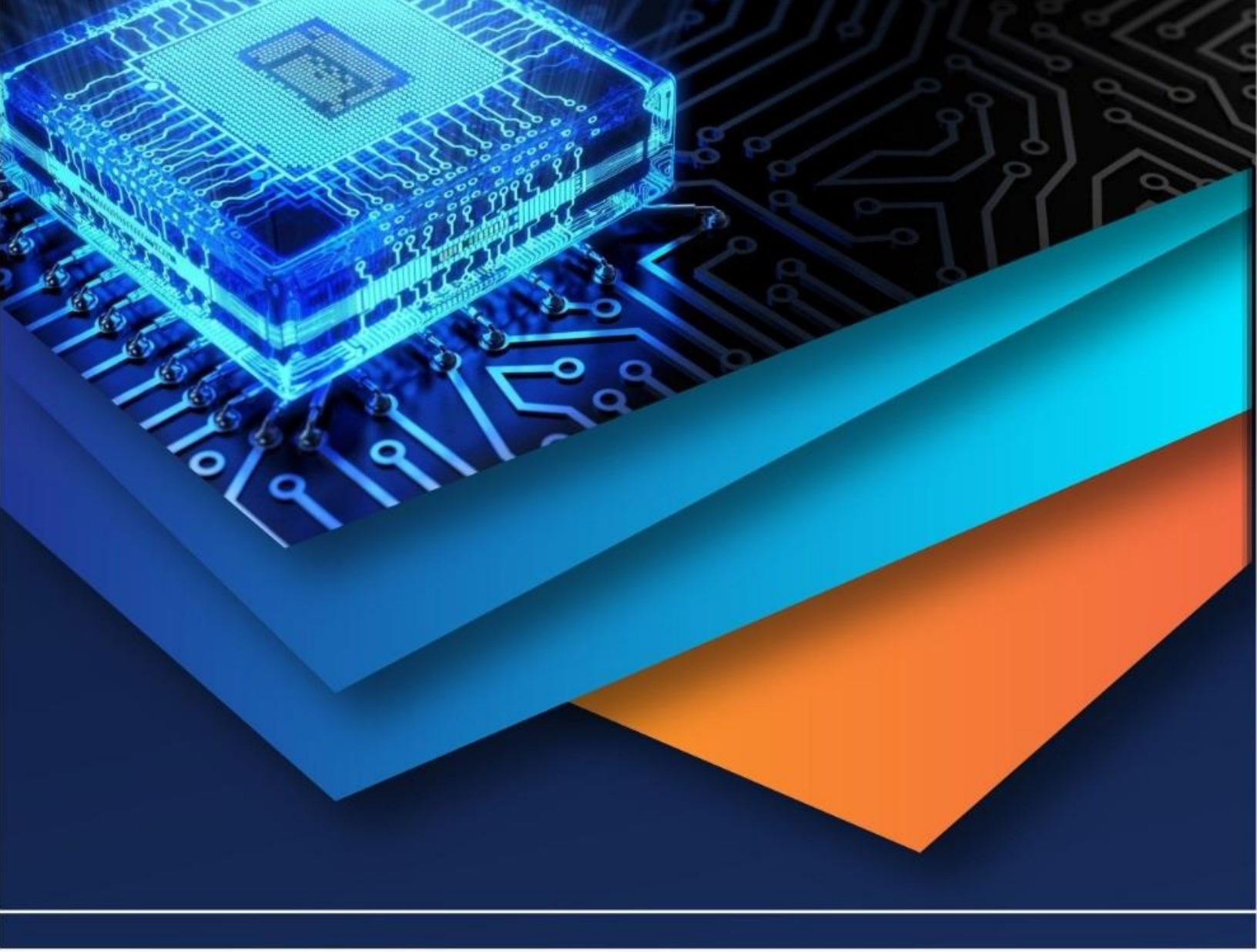
Although CyberSuraksha functions effectively as a prototype, the current implementation lacks a backend server and database, relying on LocalStorage for temporary data management. It does not yet support real-time video stream analysis. Future development will include backend integration using Firebase or PostgreSQL, implementation of a CI/CD pipeline, live-stream analysis, and dashboards for institutional monitoring and reporting.

IX. CONCLUSION

CyberSuraksha provides a practical, user-centered solution to the increasing threat of deepfakes. By combining multimodal AI capabilities with interpretability and cybersecurity education, it empowers users to identify manipulated content while developing safer online practices. The system presents a scalable framework for future research and real-world deployment, highlighting the significance of multimodal AI in digital safety.

REFERENCES

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Information Fusion, vol. 64, pp. 131–148, 2020.
[2] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE International Conference on Computer Vision, 2019.
[3] Google AI, "Gemini API Documentation," 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY