JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, P.O. – WAKNAGHAT,
TEHSIL – KANDAGHAT, DISTRICT – SOLAN (H.P.)
PIN – 173234 (INDIA) Phone Number- +91-1792-257999
(Established by H.P. State Legislature vide Act No. 14 of 2002)

Department of Computer Science & Engineering and Information Technology

Major Project Proposal (2024-25)

Group No.	64

1. Project Title

Voice Shield: AI-Powered Deepfake Audio Defense

2. Team Members

S. No.	Roll No.	Name	Mobile No.	Proficiency
1.	211293	Isha	8091104896	Machine Learning
2.	211326	Aditi Pandey	8217461804	Data Science
3.	211204	Tanisha	8800574181	Information Security

3. Name of Supervisor (s)

Dr. Ruchi Verma - Assistant Professor (SG) – CSE Department

4. Work Distribution

S. No.	Roll No.	Work Distribution
1.	211204	● Design the overall security architecture for the project. ● Ensuring data protection and secure handling of audio files. ● Oversee the secure deployment of the system. ● Ensuring that servers, databases, and APIs are hardened against attacks.
2.	211293	● Develop fine-tune machine learning models for detecting DeepFake audio. ● Use of models like CNNs or RNNs. ● Train the models using datasets of both real and DeepFake audio. ● Iterate on improving accuracy.
3.	211326	● Collect accurate datasets of real and DeepFake audio ● Handle data cleaning, normalization, and augmentation as needed.

		● Perform EDA to understand the characteristics of the data, visualize patterns, and identify anomalies. ● Analyze the results from the ML models, providing insights and recommendations for improvement.

5. Problem Statement

Fake media, generated by methods such as deepfakes, have become indistinguishable from real media, but their detection has not improved at the same pace. Furthermore, the absence of interpretability on deepfake detection models makes their reliability questionable. In this project, we present a human perception level of interpretability for deepfake audio detection. The challenge is compounded by the fact that many current deepfake detection models, particularly those designed for audio, lack interpretability. This means that while they may be able to flag a piece of audio as fake, they do not provide clear reasoning or explanations for their decisions. This lack of transparency can undermine the reliability and trustworthiness of these models, particularly in high-stakes situations where the consequences of incorrect identification could be severe. To address these issues, there is a pressing need for the development of more effective and interpretable deepfake audio detection methods. This project aims to bridge this gap by leveraging artificial intelligence (AI) techniques traditionally used in image classification and applying them to the domain of audio deepfake detection.

By drawing parallels between the cognitive processes humans use to interpret images and those used to analyze audio, this project proposes a novel approach to providing interpretability in deepfake audio detection. Specifically, we will implement several AI methods that have been successful in the visual domain and adapt them to work with audio data. Additionally, by using a corresponding data format that aligns with human cognitive processes, we aim to enhance the interpretability of the detection model, making it easier for users to understand why a particular piece of audio has been flagged as fake. Ultimately, this project seeks to create a more reliable and transparent defense against the growing threat of audio deepfakes, thereby contributing to the broader effort to safeguard digital integrity in the face of increasingly sophisticated AI-generated forgeries.

6. Main Objectives

1) **Develop an Effective Deepfake Audio Detection Model**: The primary objective of this project is to create a robust deepfake audio detection model that can accurately identify fake audio clips generated using advanced deep learning techniques.

2) **Incorporate Explainable Artificial Intelligence (XAI) Techniques**: To enhance the reliability and transparency of the detection model, this project aims to integrate explainable AI methods traditionally used in image classification into the audio domain.

3) **Enhance User Trust and Model Transparency**: A key objective is to improve user trust in the deepfake detection system by making the detection process more transparent and interpretable. By providing attribution scores and other interpretability measures, the project seeks to ensure that users can confidently rely on the system's outputs, especially in situations where accurate detection is critical.

7. Resources Required

Category	Description	
Software Resources	● Python	Version: 3.8+
	● TensorFlow/PyTorch	Version: Latest
	● Librosa	Version: Latest
	● Numpy	Version: Latest
	● Scikit-learn	Version: Latest
	● Git	Version: Latest
	● IDE	Version: Latest
Hardware Resources	● High-Performance CPU ● GPU (NVIDIA) ● 16GB RAM ● 500GB SSD	
Others	● Access to Dataset	

8. **Project Plan**

Activity	Year 2024										Year 2025									
	Aug.		Sept.		Oct.		Nov.		Dec.		Jan.		Feb.		Mar.		Apr.		May	
Literature Review	🟨	🟨	🟨	🟨																
Analysis and Requirements	🟧	🟧	🟧	🟧																
Project Design and Architecture			🟦	🟦	🟦	🟦														
Implementation					🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩						
Testing and Validation									🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧				
Documentation and Write-up									🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

Signatures

(Isha) (Aditi Pandey) (Tanisha)

(Dr. Ruchi Verma)

Date of Submission: 21st August 2024