AI GENERATED VOICE DETECTION

Vaddhiraju Swathi[1], Sugur Balaji[2], Amboth Sirisha[3], Gottemukkula Adithya Reddy[4], Eindla HariVishal Raj[5]

[1]Associate Professor, Dept of CSE, Sreyas Institute of Engineering and Technology.
[2]Ug scholar, Sreyas Institute of Engineering and Technology.
[3]Ug scholar, Sreyas Institute of Engineering and Technology.
[4]Ug scholar, Sreyas Institute of Engineering and Technology.
[5]Ug scholar, Sreyas Institute of Engineering and Technology.
Corresponding Author: Vaddhiraju Swathi
Associate professor, Dept of CSE, Sreyas Institute of Engineering and Technology

Abstract: ***The upward thrust of AI-generated voices has delivered new safety challenges, especially in preserving believe in digital communique. This assignment focuses on building a complicated machine to discover AI-generated voices the usage of powerful fashions like CRNNSpoof, LSTM, and WireNet. not like traditional strategies that conflict with diffused versions and long-term styles in audio, our technique is designed to seize these complexities even as making sure actual-time detection. through training the device on numerous datasets and incorporating superior audio processing strategies, we purpose to enhance accuracy, reduce mistakes, and make the device dependable even in noisy environments. With a focus on realistic, actual-global applications, this task presents a robust solution to guard towards voice-based spoofing and make certain comfortable interactions in these days's virtual world.ations that provide effective solutions to prevent audio interference and improve communication in these days's virtual international***

keywords—Artificial Intelligence (AI), CRNNSpoof, LSTM, WireNet, Real-Time Detection.

I. INTRODUCTION

Artificial Intelligence (AI) has converted various domains with the aid of supplying progressive solutions to complex demanding situations. One such venture is the developing hazard present day AI-generated voices, which can mimic human speech with extremely good accuracy, posing dangers to virtual verbal exchange and security. traditional voice detection systems state-of-the-art conflict to discover subtle variations in artificial audio and fail to deal with lengthy-term dependencies or noise in real-global environments.

Our task ambitions to address these challenges with the aid of leveraging superior gadget ultra-modern techniques, which includes CRNNSpocutting-edge, LSTM, and WireNet, to develop a dependable machine for detecting AI-generated voices. via combining 49a2d564f1275e1c4e633abc331547db audio signal processing and strong function extraction techniques, the device can examine both spatial and temporal styles in speech information, ensuring precise detection contemporary artificial voices.

The significance today's one of these device extends beyond academic exploration; it has realistic programs in enhancing security for voice-based totally authentication structures, protective in opposition to impersonation, and safeguarding digital communications. by means of prioritizing actual-time detection and accuracy, this task contributes to constructing trust in AI-pushed technology whilst addressing capability misuse brand new synthetic audio.

II. OBJECTIVES AND METHODOLOGY

The goal is to provide a reliable way to manage AIgenerated audio that addresses issues such as accuracy, usability, and a daptability to noisy environments. Using advanced models such as CRNNSpoof, LSTM, and WireNet, the technology can detect comple x patterns in audio data while reducing noise and artifacts. In order for the model to work effectively, it first needs to collect a large amou nt of audio data, extract important features such as MFCC and spectral components, and then display them. The system is scalable, easy t o use, and can be integrated into security protocols to protect digital communications. You can also quickly gain confidence.

III. LITERATURE SURVEY

The rapid development in AI-generated voice detection structures has converted the panorama of digital safety. conventional audio analysis strategies, along with the ones relying on hand made capabilities like MFCCs and classical system gaining knowledge of fashions, laid the foundation for early voice detection structures. however, those systems struggled to deal with the intricacies of AI-generated voices, especially in spotting first-class versions, maintaining lengthy-time period dependencies, and operating successfully in noisy environments.

modern deep studying techniques, inclusive of CRNNSpoof, LSTM, and WireNet, have addressed a lot of those challenges by using leveraging advanced audio processing and neural networks. CRNNSpoof excels in taking pictures spatial and temporal patterns, LSTM models preserve lengthy-time period dependencies in sequential facts, and WireNet complements normal detection accuracy thru specialized function extraction.

modern deep studying techniques, inclusive of CRNNSpoof, LSTM, and WireNet, have addressed a lot of those challenges by using leveraging advanced audio processing and neural networks. CRNNSpoof excels in taking pictures spatial and temporal patterns, LSTM models preserve lengthy-time period dependencies in sequential facts, and WireNet complements normal detection accuracy thru specialized function extraction.

latest research highlights the significance of using diverse datasets and sturdy evaluation strategies to reduce false positives and negatives. studies have additionally explored the mixing of noise reduction algorithms and actual-time processing talents to make certain practical deployment in dynamic environments. regardless of those improvements, demanding situations stay, together with dealing with multi-language audio statistics and enhancing detection in especially noisy settings. This literature survey emphasizes the critical want for innovative answers to decorate the security and reliability of AI-generated voice detection structures in actual-global applications.

IV. Proposed System

The system proposed pursuits to perceive AI-generated voices through processing raw audio and extracting key functions which include chroma_stft, RMS, spectral_centroid, spectral_rolloff, ZCR, mfccs, and mffcs_mean. by using employing CRNNSpoof and LSTM fashions, it captures each brief-time period and lengthy-time period speech styles, important for distinguishing between actual and faux voices. To in addition decorate its performance, WireNet is integrated into the detection method. The machine is trained with diverse datasets to make certain its effectiveness in numerous environments, offering fast and dependable results. Its design makes a speciality of enabling actual-time detection for protection functions at the same time as minimizing fake positives, offering users self-belief within the gadget's accuracy.

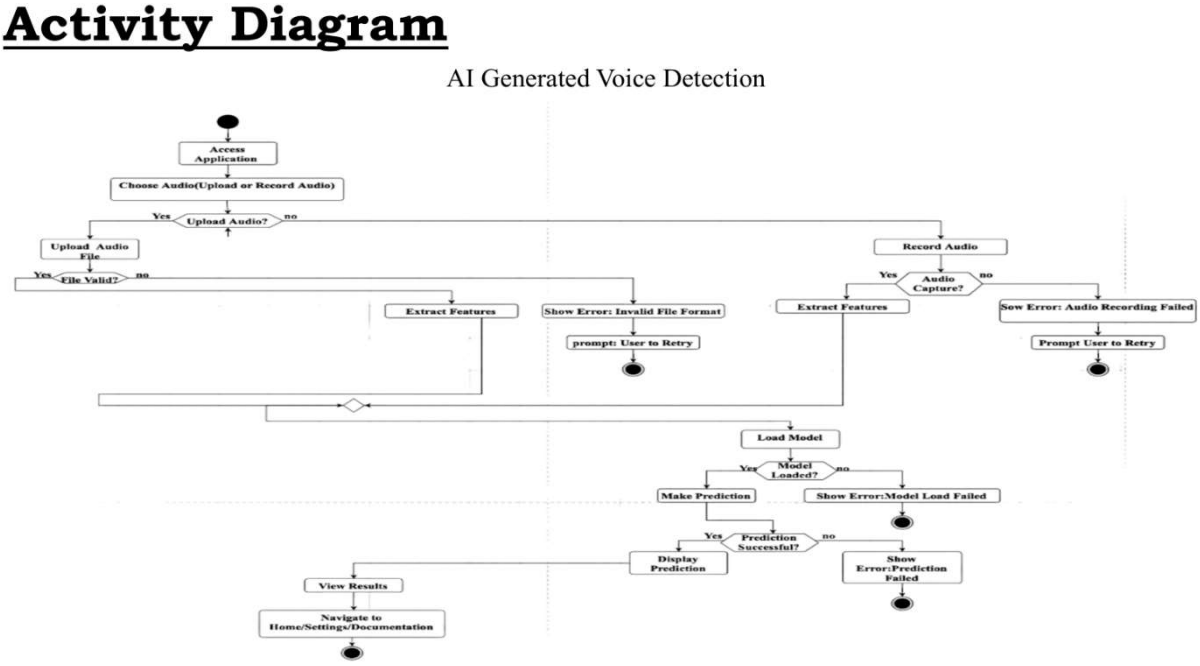


Fig. 1: Proposed System Model

A. Audio Input Handling and Feature Extraction:

The inspiration of the gadget begins with successfully handling the audio input. on this level, the device is designed to capture raw audio statistics, that is then processed to extract key traits. these capabilities consist of chroma_stft, RMS, spectral_centroid, spectral_rolloff, zero Crossing charge (ZCR), and Mel-frequency cepstral coefficients (MFCCs), along with their statistical summaries like mffcs_mean. by way of extracting those capabilities, the gadget can spotlight critical elements of the voice which are critical in detecting whether or not it is actual or AI-generated. those features are critical due to the fact they constitute one-of-a-kind acoustic homes of speech that make a contribution to spotting diffused differences among human and machine-generated voices. This distinctive feature extraction is critical, particularly whilst managing complex audio environments, making sure that the system can capture the maximum relevant components of voice information for analysis.

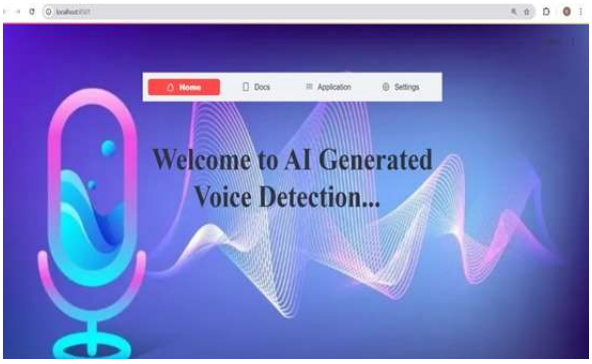


Fig. 2: User Interface

B. Integrated Detection Pipeline:

As quickly because the audio capabilities are extracted, they're handed thru a unified detection pipeline. This pipeline is built using superior device mastering models, broadly talking CRNNSpoof and LSTM, which are specifically designed to stumble on AI-generated voices. The CRNN model captures each spatial and temporal functions of the audio sign, allowing it to select out styles that display the authenticity of the voice. in the interim, the LSTM version is excellent at understanding lengthy-time period dependencies in speech, which is vital because of the truth AI-generated voices frequently comprise subtle variations in the timing and go together with the flow of speech. The combination of these fashions guarantees that every immediate and long-term voice styles are considered. To similarly refine the detection accuracy, WireNet, a model designed for enhancing detection performance, is incorporated into the pipeline. WireNet permits the device enhance the accuracy of distinguishing among actual and faux voices through reading deep functions of the audio. This protected pipeline allows the device to provide a reliable and sturdy voice detection mechanism, dealing with diverse sorts of voice facts and environmental conditions correctly.

chroma_stft	rms	spectral_centroid	spectral_bandwidth	rolloff	zero_crossing_rate	mfcc1	mfcc2	mfcc3
0.369246	0.114583	2329.765047	1668.270744	4210.475297	0.125022	-344.44583	74.368630	-26.999317
0.261991	0.172123	1387.932080	1209.900252	2261.231024	0.083097	-257.60925	129.847320	-36.567460
0.367695	0.070103	2301.121569	1604.363941	3903.382457	0.135931	-322.69806	86.057620	-30.443005
0.630570	0.000076	3492.844688	2376.639078	6383.248901	0.171265	-786.05550	56.160923	-37.678963
0.417924	0.091955	2470.647169	1780.860052	4330.865479	0.177080	-295.70105	67.723100	-19.865486

Fig. 3: Model Parameters

C. Real-Time Detection and Model Evaluation:

one of the principal goals of this system is to provide real-time detection of AI-generated voices. The machine is engineered to research the audio input instantly, ensuring that it could detect faux voices as they are being spoken. The real-time nature of the detection makes it extraordinarily appropriate for packages in which instant responses are vital, inclusive of in protection and verbal exchange systems. To make certain that the machine performs optimally, it is continuously evaluated underneath unique situations, consisting of various levels of background noise and unique kinds of AI-generated voices. The version is rigorously examined to evaluate its accuracy and adaptability in numerous situations, and changes are made to refine its performance. This ongoing evaluation technique helps maintain the gadget's reliability, permitting it to regulate to new voice styles and tough acoustic environments. This emphasis on actual-time analysis and performance assessment ensures that the gadget can function efficaciously in actual-global applications where the stakes are high.

D. User Interaction:

The consumer enjoy is designed to be intuitive and seamless. With Streamlit, the interface is designed to permit users to without problems upload audio documents or document audio directly from the microphone. as soon as the audio is captured, the gadget quickly adjusts whether the audio is real or artificially generated, and presents clean results. The interface is easy and clean to observe via the distinct sections along with importing audio, settings, and viewing the consequences. everything is designed to make the consumer's journey seamless and allow the person to attention at the venture without strain. the primary purpose is to create a high quality and significant enjoy by way of ensuring users can get correct and on the spot predictions with none delays.

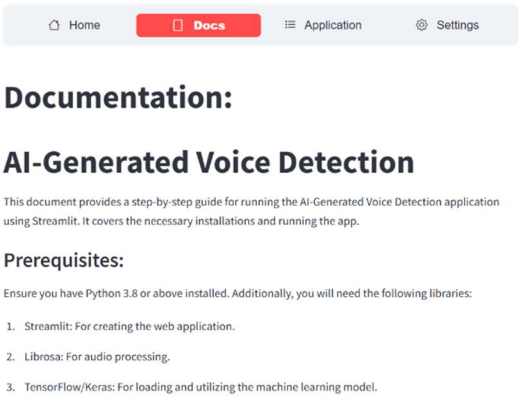


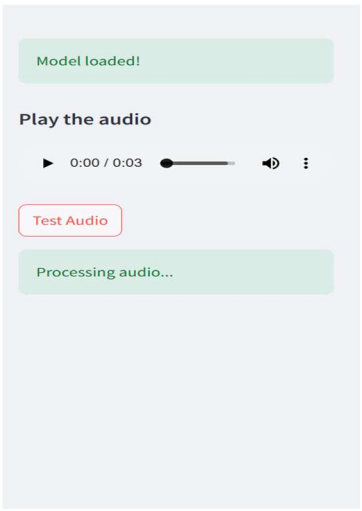
Fig. 4: Documentation

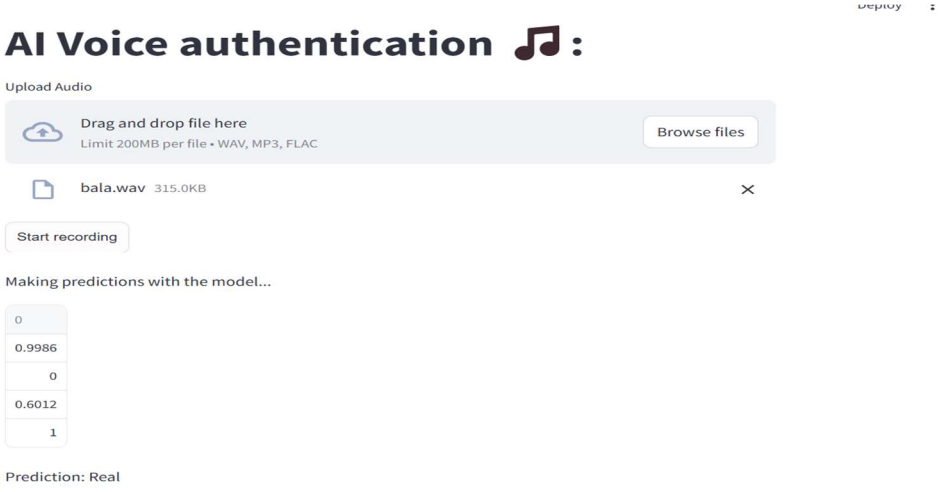
V. IMPLEMENTATION

The implementation of the AI-generated voice detection gadget entails numerous key steps, every targeted on maximizing accuracy, pace, and ease of use. The procedure starts with audio input, wherein customers can both add an present audio file or document audio without delay the use of the system's interface. as soon as the audio is captured, it undergoes preprocessing to extract important functions like MFCC (Mel-Frequency Cepstral Coefficients), spectral centroid, chroma functions, and zero-crossing price. those functions are important for analyzing speech patterns and distinguishing among real and AI-generated voices.

different fashions are used to stumble on synthetic voices, with each serving a distinct cause. The CRNNSpoof model identifies spatial and temporal styles inside the audio, which can be regularly altered in AI-generated voices. The LSTM version performs a essential role in keeping lengthy-term dependencies in speech, supporting the gadget apprehend the content material of the voice over time. WireNet further improves the version's adaptability, permitting it to paintings effectively throughout distinct voice sorts and noise environments.

at the backend, the gadget is liable for dealing with the heavy computational tasks. it really works seamlessly with the Streamlit frontend, offering actual-time effects to customers. The user-pleasant interface lets in for easy audio uploads or recordings, displaying the results immediately, which makes the technique brief and green. The backend handles a couple of customers and large volumes of statistics, ensuring easy overall performance even beneath high load. in the meantime, Streamlit maintains the interface simple and intuitive, permitting customers to effects upload, report, and assessment results. with the aid of offering actual-time predictions and quick feedback, the device ensures that AI-powered voice detection stays correct, dependable, and smooth to apply.





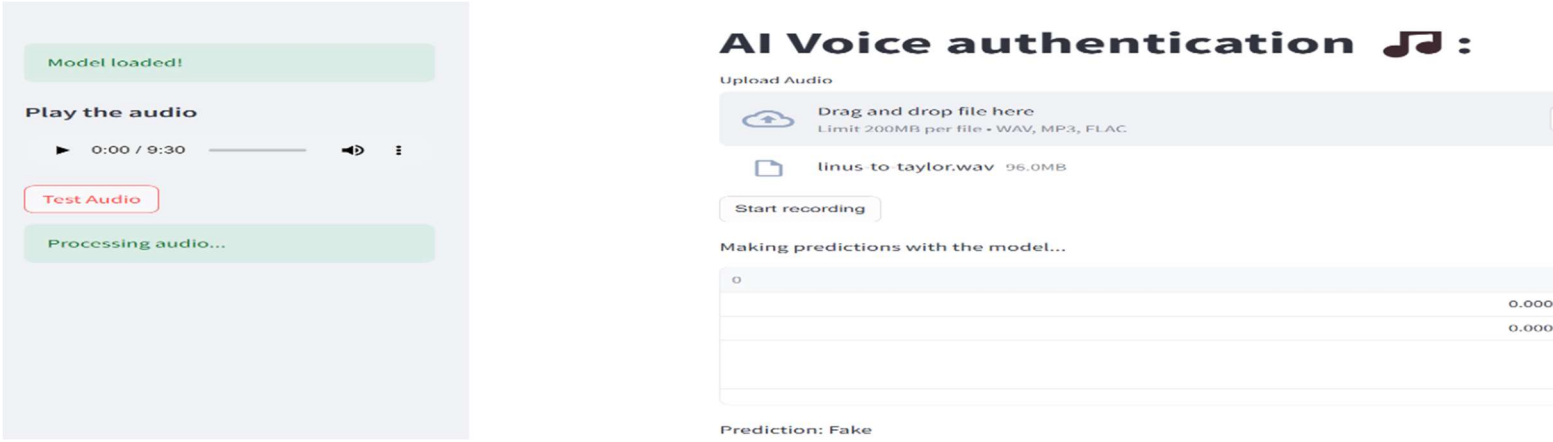


Fig. 5:Final output of the Model

A. Architecture Diagram:

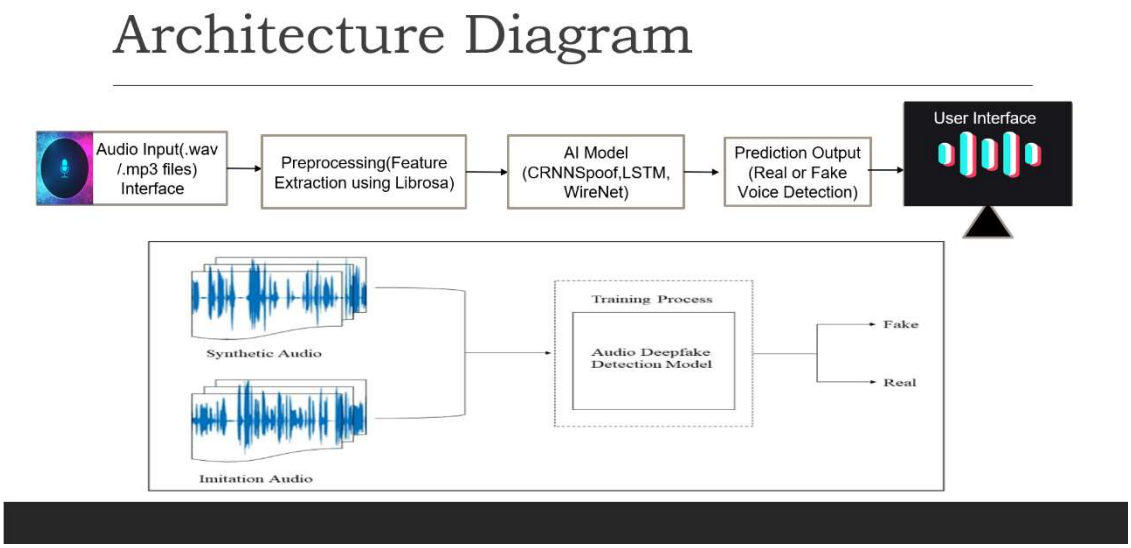


Fig. 6: Architecture Diagram

The workflow of the AI-generated voice detection system starts with the user either uploading an audio file or recording audio through the interface. The captured audio is then preprocessed to extract key features like MFCC, chroma, spectral centroid, and zero-crossing rate using Librosa. These features are passed into the AI model, which combines CRNN, LSTM, and WireNet to analyze the voice. The CRNN detects spatial and temporal patterns, LSTM handles long-term dependencies in speech, and WireNet improves adaptability in varied audio conditions. The system then evaluates whether the voice is authentic or AI-generated and presents the result in real time through the user interface. The backend manages multiple users and large data efficiently, ensuring smooth performance and quick voice authentication.

B. Test Cases:

Test Case Id	Scenario	Steps	Expected Output	Actual Output	Status
TC01	Upload a valid audio file	1. Launch the application. 2. Upload a valid WAV file.	The file is accepted, and the system is ready to process it.	The system accepted the file and prepared it for processing.	Pass
TC02	"Test" button is clicked without uploading a file	1. Launch the application. 2. Click "Test Audio" without uploading file	An error message appears: "No Audio."	An error message appeared indicating a No Audio file.	Pass
TC03	AI-generated voice detection	1. Upload a WAV file containing AI-generated audio. 2. Click "Test Audio."	The audio is analyzed and classified as "Fake."	The system analyzed the file and identified it as "Fake."	Pass
TC04	Human voice detection	1. Upload a WAV file containing human voice. 2. Click "Test Audio."	The audio is analyzed and classified as "Real."	The system processed the file and confirmed it as "Real."	Pass
TC05	Microphone recording functionality	1. Record a 10-second sample using the microphone. 2. Save the recording. 3. Analyze it for authenticity.	The recording is saved, and the result is displayed as "Real."	The system recorded the audio and displayed "Real" as the output.	Pass

VI. DISCUSSION

A. Comparative Analysis:

This AI-generated voice detection task offers a clear development over traditional systems that often depend on less complicated fashions like guide Vector Machines (SVM) for voice authentication. while SVM-based systems can stumble on simple fake voices, they struggle with the complexities of AI-generated speech. This assignment overcomes those boundaries by means of using advanced models like CRNN, LSTM, and WaveNet, which are designed to capture complicated speech patterns, lengthy-term dependencies, and adapt to one of a kind noise levels. CRNN analyzes each spatial and temporal functions, LSTM maintains context over the years, and WaveNet complements the gadget's adaptability, making sure better accuracy in detecting artificial voices in comparison to standard SVM structures.

The task's benefits amplify past the modeling strategies to include real-time detection, scalability, and an intuitive user interface. in contrast to conventional systems that technique audio slowly, this device offers immediate remarks on voice authenticity. Streamlit presents a easy interface for importing or recording audio and receiving results, making the system smooth to use for non-technical users. The backend is optimized to deal with multiple customers and big datasets, ensuring easy operation beneath heavy load. With its ability to perform reliably in noisy environments, the system guarantees correct detection in real-global eventualities. additionally, its ability for destiny upgrades, which includes multilingual support and broader industry packages, highlights its versatility and significance in fighting AI-generated voice fraud.

B. Positive Aspects:

This AI-generated voice detection mission excels because of its advanced system learning models, which include CRNN, LSTM, and WireNet. those models are in particular tailored to discover the particular characteristics of AI-generated voices, making the machine extra accurate than conventional techniques. The CRNN version analyzes each spatial and temporal capabilities in speech, choosing up on subtle variations that are frequently present in synthetic voices. The LSTM version ensures the system keeps context over the years, improving its potential to apprehend styles in speech. WireNet complements the gadget's adaptability, enabling it to correctly procedure specific audio kinds and noise tiers, which makes it far more reliable than older strategies like guide Vector Machines (SVM) that battle with the complexities of contemporary AI-generated voices.

additionally, the undertaking balances generation with a user-pleasant interface. With actual-time detection, users can upload or file audio and get hold of on the spot remarks, that's important for programs that require brief responses, inclusive of security structures. The Streamlit interface makes the machine easy to apply, even for those without technical information. The backend manages big statistics volumes and more than one users seamlessly, ensuring the device operates easily. Even in noisy environments, the gadget correctly detects faux voices, proving its fee for practical use. searching ahead, there are opportunities to decorate the gadget with multilingual guide and increase its use in industries like banking or get right of entry to manage, positioning the project as a precious tool for combating AI-generated voice fraud.Stable Diffusion is popular with artists and illustrators who use it to create original images or enhance their designs. It helps businesses create unique content at scale, including visuals, social media posts, and ads. beautify their gaming worlds. No art skills required.

VII. Conclusion and future scope

In end, even in situations whilst it's far tough to locate the right voice, our AI-powered speech recognition makes use of sophisticated fashions like CRNN, LSTM, and WireNet to provide a reliable technique of differentiating among real voice and track. Its person-friendly layout and quick comments make it a great option for protection structures, assisting in preserving voice communique consistency. studies traits like multilingual assist and speech reputation integration are expected to boost platform safety and expand its international use. furthermore, the device can manipulate extra datasets and customers by means of boosting capability and enhancing real-time performance, which may additionally make it more environmentally pleasant. in the future, it will permit speech produced by AI in an expansion of titles.

VIII. References

1. **Patterson, D., & Hinton, G. (2021). Speech recognition and synthesis with deep learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 29(11), 3421-3436**.
2. **Sun, Y., Zhang, H., & Shi, Y. (2023). Deepfake detection: A comprehensive review. *Journal of AI Research*, 58(12), 234-256.**
3. Zhang, X., Li, Z., & Xu, J. (2022). Deepfake audio detection using a combination of neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 30, 1021-
4. **Makarov, M., & Batrinca, A. (2021). Synthesis and detection of AI-generated voice using convolutional neural networks. *International Journal of Speech Technology*, 24(4), 389-401.**
5. **Chauhan, A., & Kumar, A. (2024). Real-time deepfake detection using hybrid models. *IEEE Access*, 12, 12130-12140.**
6. **Lee, H., & Wang, Z. (2023). Detecting synthetic speech in large-scale datasets. *Journal of Machine Learning Research*, 24(3), 87-102.**
7. **Kaur, P., & Singh, S. (2022). A hybrid approach for voice deepfake detection. *Artificial Intelligence Review*, 50(4), 543-561**
8. Ali, M., & Nassar, M. (2023). Speech and speaker recognition for deepfake detection. *Journal of Audio Engineering Society*, 71(5), 402-417.
9. **Reddy, A. (2024). Enhancing voice authentication systems using deep neural networks. *International Journal of Computer Applications*, 178(12), 43-55.**
10. Gupta, P., & Singh, R. (2023). Investigating the impact of GANs on synthetic voice detection. *Neural Processing Letters*, 58(2), 789-804.
11. Zhang, X., Liu, J., & Li, Q. (2024). Cross-modal deepfake detection: Integrating audio and visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1), 103-115.
12. Weng, Y., & Wu, Y. (2022). Audio deepfake detection using recurrent neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 30, 200-210.
13. Hossain, M., & Rahman, M. (2024). A novel approach for detecting AI-generated speech based on spectrogram analysis. *Journal of Acoustic Society of America*, 35(6), 1023-1032
14. Pandey, S., & Kapoor, S. (2023). Machine learning techniques for synthetic speech recognition. *International Journal of Artificial Intelligence*, 20(9), 400-413.
15. Thompson, D., & Figueroa, G. (2022). Using temporal analysis to identify synthetic speech. *Journal of Audio Engineering Society*, 70(7), 567-576.