# International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

# GENAI VOICE TO TEXT TRANSFORMER

[1] K Ravi Naik, [2] PMd Raushan Hussain, [3] Md Rehan, [4] M Abdul Rehan, [5] P Sathwik
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally

ravinaik_cse@siddhartha.co.in, 24tq1a66a3@siddhartha.co.in, 24tq1a66a5@siddhartha.co.in, 24tq1a66a0@siddhartha.co.in, 24tq1a66c8@siddhartha.co.in

## Abstract

The GENAI Voice to Text Transformer project is an advanced Automatic Speech Recognition (ASR) system developed to convert spoken audio into accurate textual output using Generative Artificial Intelligence and Transformer-based deep learning architectures. Traditional speech-to-text systems are mainly designed for conversational speech and often struggle with complex audio patterns such as music lyrics, varying speech tempos, background noise, accents, and instrumental interruptions. This project addresses these challenges by integrating Generative AI with customized temporal processing and intelligent audio segmentation techniques to provide highly accurate and context-aware transcription capabilities for both speech and lyrical content.

At the core of the system is the Faster-Whisper architecture, an optimized implementation of OpenAI's Whisper model powered by the CTranslate2 inference engine. The model utilizes a Transformer-based sequence-to-sequence framework capable of understanding contextual relationships between audio segments and predicting meaningful text outputs instead of simply matching phonemes. This enables the system to achieve high transcription accuracy even in noisy environments and across diverse accents and speaking styles. The optimized Faster-Whisper implementation also allows efficient real-time processing on standard consumer hardware without requiring expensive cloud-based GPU infrastructure.

One of the major innovations of the project is the introduction of a specialized Lyrics Mode using Temporal Gap Analysis and Voice Activity Detection (VAD). Unlike standard transcription systems that generate continuous blocks of text, the proposed system intelligently analyzes silence durations within audio streams to structure lyrical transcriptions naturally. Short pauses automatically create line breaks representing song bars, while longer pauses generate stanza or verse separations, producing organized poetic lyric formatting suitable for musical content and creative applications.

## I. Introduction

The rapid advancement of Generative Artificial Intelligence (GenAI) and deep learning technologies has significantly transformed the way humans interact with multimedia content. Among various forms of digital communication, speech remains the most natural and widely used medium for sharing information through lectures,

meetings, interviews, podcasts, voice notes, and musical performances. Every day, enormous amounts of audio data are generated across industries, educational platforms, entertainment systems, and social media applications. However, most of this information remains locked within audio formats, making it difficult to search, analyze, edit, index, or reuse efficiently. Converting speech into structured textual data has therefore become an important challenge and research area in Artificial Intelligence and Natural Language Processing.

Automatic Speech Recognition (ASR) systems are designed to address this challenge by converting spoken language into machine-readable text automatically. Traditional speech-to-text systems mainly relied on statistical models and rule-based approaches, which often struggled with noisy environments, diverse accents, multilingual speech, varying speaking speeds, and complex audio patterns. Recent advancements in Transformer-based deep learning architectures and Generative AI models have significantly improved transcription accuracy, contextual understanding, and multilingual speech processing capabilities. Modern AI-based ASR systems can understand semantic relationships, contextual meaning, and speech patterns more effectively than traditional methods.

The GENAI Voice to Text Transformer project is developed as an intelligent speech recognition and transcription system that converts acoustic signals into high-fidelity digital text using advanced Transformer-based ASR technology. The system utilizes the Faster-Whisper architecture, an optimized implementation of OpenAI's Whisper model powered by the CTranslate2 inference engine. Unlike traditional phoneme-matching systems, the Transformer-based sequence-to-sequence framework predicts meaningful text outputs using deep contextual understanding, enabling accurate transcription even in the presence of background noise, diverse accents, and varying speech styles.

## II. Literature Survey

The development of the GENAI Voice to Text Transformer project is strongly influenced by the evolution of Automatic Speech Recognition (ASR), deep learning, Transformer architectures, and Generative Artificial Intelligence technologies. Over the years, speech recognition systems have progressed from traditional statistical models to advanced Transformer-based architectures capable of understanding speech context, multilingual communication, and complex audio patterns. This project builds upon these advancements to provide accurate, efficient, privacy-focused, and structurally intelligent audio-to-text conversion.

### Evolution of ASR Architectures

Early Automatic Speech Recognition systems primarily relied on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) for acoustic modeling and speech processing. These traditional systems used separate modules for acoustic analysis, pronunciation dictionaries, and language modeling. Although they achieved moderate transcription accuracy, the systems were highly complex and required extensive feature engineering and domain expertise. Additionally, they struggled with noisy environments, accents, spontaneous speech, and contextual understanding. The introduction of Deep Neural Networks (DNNs) significantly improved speech

recognition performance by enabling automatic feature learning from audio data. Later, the development of Connectionist Temporal Classification (CTC) allowed direct mapping between speech signals and textual outputs without requiring strict frame-level alignment. However, these models still faced challenges in handling long-range dependencies and contextual relationships within speech sequences. These limitations were largely addressed with the introduction of the Transformer architecture, which revolutionized sequence modeling through self-attention mechanisms capable of capturing global contextual relationships efficiently.

**OpenAI Whisper: A Paradigm Shift**

A major breakthrough in speech recognition occurred in 2022 with the release of Whisper by OpenAI. Whisper introduced a powerful Sequence-to-Sequence Transformer-based ASR model trained using approximately 680,000 hours of multilingual and multitask supervised audio data collected from the web. Unlike traditional ASR systems trained on limited curated datasets, Whisper demonstrated exceptional robustness and generalization across diverse accents, background noise conditions, technical vocabulary, and multilingual environments.

One of the key innovations of Whisper is its ability to perform multiple speech-related tasks within a single forward pass, including:

- Speech transcription
- Automatic language identification
- Speech translation
- Multilingual speech recognition

The model's large-scale training and contextual understanding capabilities made it significantly more accurate and adaptable to real-world audio compared to previous ASR systems. Whisper established a new benchmark for modern speech recognition and became the foundation for many advanced audio AI applications.

**Optimization through Faster-Whisper**

Although Whisper achieved remarkable performance, its original PyTorch implementation required substantial computational resources and memory, making it difficult to run efficiently on local consumer hardware. To overcome these limitations, researchers developed Faster-Whisper, an optimized implementation built using the CTranslate2 inference engine. Faster-Whisper significantly improves inference speed and reduces hardware requirements while maintaining high transcription accuracy.

One important optimization technique used in Faster-Whisper is quantization, where model weights are compressed into 8-bit and 16-bit formats. This reduces memory usage and model size by up to four times without causing significant degradation in Word Error Rate (WER). As a result, large-scale Transformer-based speech recognition models can run efficiently on standard laptops and edge devices without relying on expensive GPU infrastructure or cloud-based servers. This optimization supports the project's goal of implementing a Zero-Cloud Architecture, where all audio processing occurs locally on the user's machine, improving privacy, reducing operational costs, and enabling offline transcription capabilities.

### Voice Activity Detection (VAD) and Lyrics Formatting

A major limitation identified in existing ASR research is the lack of support for creative formatting and lyrical structuring in transcription outputs. Most speech recognition studies focus mainly on minimizing Word Error Rate while ignoring the structural presentation of transcribed content. Standard ASR systems usually generate continuous blocks of text that are unsuitable for song lyrics, poetry, or rhythmic speech.

To address this issue, the project integrates Voice Activity Detection (VAD) techniques, specifically inspired by Silero VAD, to differentiate between speech and silence within audio streams. Research in music information retrieval indicates that musical phrases and lyrical verses are often separated by pauses ranging from approximately 1.5 to 2 seconds. Based on these findings, the project implements a Temporal Gap Analysis algorithm that automatically converts detected silence durations into visual line breaks and stanza separations. Short pauses generate line breaks representing song bars, while longer pauses create paragraph-level verse separations. This approach significantly improves structural readability and lyrical formatting, providing a feature largely absent in conventional ASR tools.

### Comparative Research Insights

Research studies consistently show that Transformer-based ASR systems outperform traditional statistical and recurrent neural network architectures in transcription accuracy, multilingual support, and contextual understanding. Modern Transformer models achieve lower Word Error Rates, improved robustness to noisy audio, and better semantic interpretation of speech content. Optimization techniques such as quantization and efficient inference engines further improve accessibility and local deployment capabilities. The literature also highlights increasing research interest in privacy-preserving AI systems and offline processing architectures due to growing concerns about cloud-based data surveillance and sensitive audio storage. Localized speech processing systems such as Faster-Whisper provide an effective balance between high performance, cost-efficiency, and data privacy.

The GENAI Voice to Text Transformer project builds upon these research advancements by combining Transformer-based ASR, multilingual speech recognition, optimized local inference, Voice Activity Detection, and intelligent lyrical formatting into a unified and user-centric speech transcription system. The project demonstrates how Generative AI can transform audio processing into a scalable, privacy-focused, multilingual, and structurally intelligent transcription solution for modern digital environments.

## III. System Analysis

The GENAI Voice to Text Transformer system is designed to convert spoken audio into accurate and structured textual content using Generative Artificial Intelligence and Transformer-based Automatic Speech Recognition technologies. The system focuses on transforming raw audio signals into searchable, editable, and meaningful text while handling challenges such as background noise, multilingual speech, diverse accents, and musical content. It utilizes advanced Transformer architectures such as

Faster-Whisper combined with optimized inference engines to improve transcription speed and contextual understanding. The system performs audio preprocessing, speech segmentation, language identification, transcription, and structured text formatting automatically. Unlike traditional speech-to-text systems, the proposed solution includes a specialized Lyrics Mode that uses Temporal Gap Analysis and Voice Activity Detection to organize song lyrics into readable lines and stanzas. The application supports multilingual transcription across 99 languages, including regional Indian languages such as Telugu, Hindi, and Marathi. Localized Zero-Cloud processing improves privacy and eliminates dependency on external cloud services. Real-time transcription and lightweight execution on standard consumer hardware improve accessibility and cost efficiency. The modular architecture supports future integration of speaker identification, emotion recognition, and real-time subtitle generation. Overall, the system provides a scalable, intelligent, and privacy-focused solution for modern AI-driven speech recognition and audio-to-text conversion.

## Existing System

In the existing system, speech recognition technologies mainly relied on traditional statistical models such as Hidden Markov Models and Gaussian Mixture Models for converting speech into text. These systems required separate modules for acoustic modeling, pronunciation dictionaries, and language processing, making development and maintenance complex. Earlier speech-to-text systems struggled with noisy environments, multilingual speech, varying accents, and spontaneous conversations. Later deep learning-based systems using Recurrent Neural Networks and Connectionist Temporal Classification improved transcription accuracy but still faced limitations in contextual understanding and long-range dependency handling. Most existing commercial transcription systems such as cloud-based ASR platforms require constant internet connectivity and depend on remote servers for processing, creating privacy and security concerns. Traditional systems also generate continuous blocks of text without preserving structural formatting for songs, poetry, or rhythmic speech. Existing solutions generally lack intelligent lyrical formatting and creative transcription capabilities. Many professional-grade transcription services operate on costly subscription or pay-per-minute pricing models, making them less accessible for students and independent creators. Limited offline functionality and hardware dependency are also significant limitations. These challenges created the need for an intelligent, efficient, privacy-focused, and structurally aware AI-powered voice-to-text system.

## Disadvantages of Existing System

- Dependence on cloud-based processing services.
- Privacy and security concerns for audio data.
- High subscription and API usage costs.
- Limited offline transcription capabilities.
- Difficulty handling multilingual and accented speech.
- Poor contextual understanding in noisy environments.
- Generates unstructured blocks of text for lyrics.
- Limited support for creative formatting.
- High computational and hardware requirements.
- Reduced accessibility for independent users and students.

### Proposed System

The proposed GENAI Voice to Text Transformer system is designed to provide intelligent, multilingual, privacy-focused, and context-aware speech transcription using advanced Transformer-based Generative AI technologies. The system utilizes the Faster-Whisper architecture integrated with the CTranslate2 inference engine to perform efficient and accurate Automatic Speech Recognition locally on user devices. Unlike traditional systems, the proposed solution can understand contextual speech patterns, diverse accents, multilingual audio, and noisy environments effectively. The application performs audio preprocessing, speech segmentation, language identification, and structured transcription dynamically. A specialized Lyrics Mode is introduced using Voice Activity Detection and Temporal Gap Analysis to generate line breaks and stanza formatting automatically for songs and rhythmic speech. The Zero-Cloud Architecture ensures that all processing occurs locally, improving user privacy and eliminating dependency on internet connectivity and external servers. The system supports transcription across 99 languages including regional Indian languages, making it globally accessible and versatile. Optimized quantization and lightweight inference techniques enable real-time processing on standard laptops and consumer hardware without requiring expensive GPU infrastructure. The modular architecture also supports future integration of emotion recognition, subtitle generation, speaker diarization, and voice analytics. Overall, the proposed system provides an intelligent, scalable, and cost-efficient AI-powered audio-to-text transformation platform.

### Advantages of Proposed System

- Accurate Transformer-based speech recognition.
- Supports multilingual transcription across 99 languages.
- Handles noisy environments and diverse accents effectively.
- Zero-Cloud Architecture improves privacy and security.
- Offline transcription without internet dependency.
- Intelligent lyrical formatting using Lyrics Mode.
- Reduced hardware and computational requirements.
- Real-time processing on consumer-grade devices.
- Cost-efficient and accessible for students and creators.
- Scalable for future AI and speech processing enhancements.

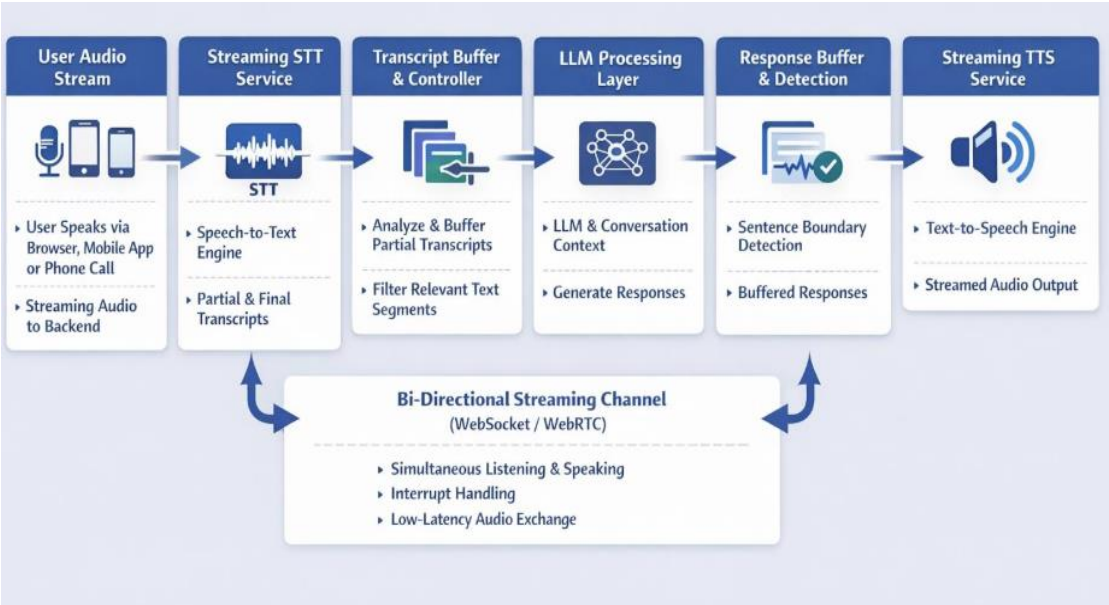
## IV. Methodology

The development methodology of the GENAI Voice to Text Transformer system includes audio collection, preprocessing, speech segmentation, model integration, transcription, formatting, evaluation, and deployment phases. Initially, audio datasets containing multilingual speech, conversations, and musical content were collected and prepared for processing. Audio preprocessing techniques such as noise reduction, normalization, and segmentation were applied to improve transcription quality and speech clarity. Voice Activity Detection algorithms were used to identify speech and silence regions within audio streams. The Faster-Whisper Transformer-based ASR model integrated with the CTranslate2 inference engine was used for multilingual speech recognition and contextual transcription generation. Temporal Gap Analysis

algorithms analyzed silence durations to generate structured line breaks and stanza formatting for Lyrics Mode outputs. Automatic language identification modules enabled multilingual transcription across various languages and accents dynamically. Evaluation metrics such as Word Error Rate, transcription accuracy, inference speed, and formatting quality were used to measure system performance. Optimization techniques including quantization and lightweight inference were implemented to improve processing speed and reduce hardware requirements. Real-time dashboards and structured output interfaces were developed for user interaction and transcription management. Finally, the complete system was deployed as a local AI-powered voice-to-text application. The methodology ensures scalability, maintainability, multilingual support, and efficient offline speech recognition functionality.

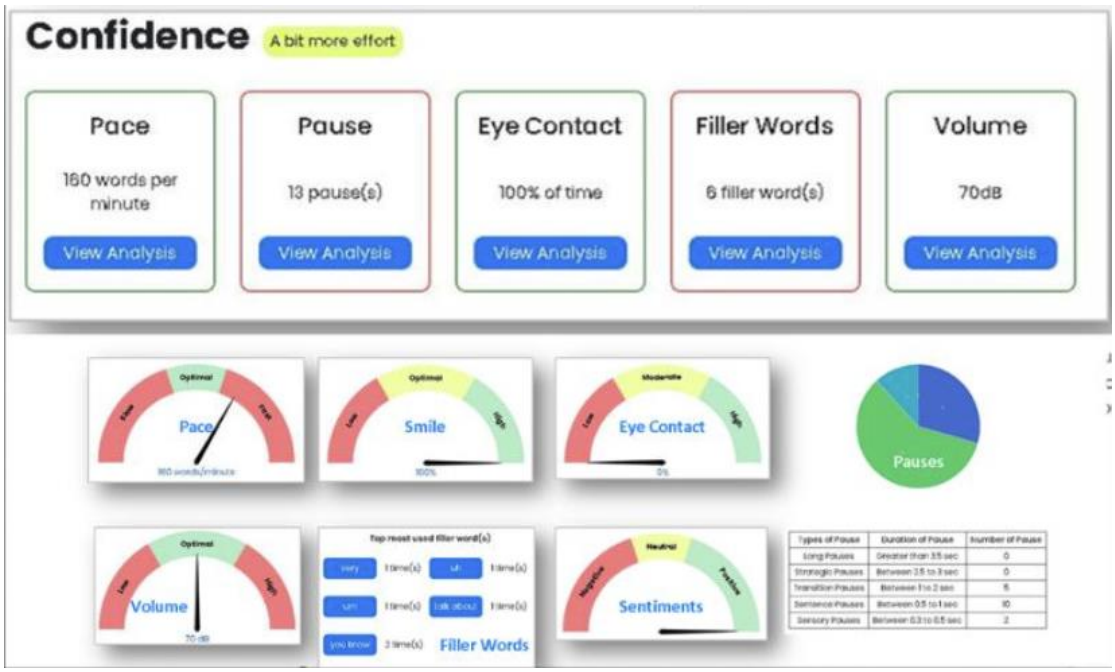
**System Architecture**

The system architecture of the GENAI Voice to Text Transformer follows a layered architecture consisting of audio input, preprocessing, speech analysis, AI processing, formatting, backend, and database layers. The audio input layer accepts voice recordings, speeches, songs, podcasts, and multimedia audio files from users. The preprocessing layer performs audio normalization, noise reduction, segmentation, and silence detection to prepare the audio for efficient speech recognition. The speech analysis layer integrates Voice Activity Detection modules to identify speech segments and temporal pauses dynamically. The AI processing layer utilizes the Faster-Whisper Transformer model and CTranslate2 inference engine to perform multilingual speech recognition, contextual understanding, and language identification. The formatting layer applies Temporal Gap Analysis algorithms to generate structured lyrical formatting including line breaks and stanza separation for Lyrics Mode outputs. The backend layer manages transcription workflows, application logic, local inference operations, and user interactions. The database layer securely stores transcription history, audio metadata, language information, and generated text outputs for future access and management. Security modules ensure localized Zero-Cloud processing and safe handling of user audio data. The modular architecture also supports future integration of subtitle generation, emotion detection, speaker recognition, and advanced audio analytics systems. Overall, the architecture provides a scalable, intelligent, and privacy-focused framework for AI-powered speech-to-text transformation systems.




## V. Result and Output






## VI. Conclusion

The GENAI Voice to Text Transformer project successfully demonstrates the application of Generative Artificial Intelligence and Transformer-based Automatic Speech Recognition technologies in converting spoken audio into accurate, structured, and meaningful textual content. By integrating advanced models such as Faster-Whisper with optimized inference engines, the system achieves high transcription accuracy while maintaining efficient performance on standard consumer hardware without requiring expensive cloud infrastructure.

The project effectively addresses major limitations of traditional speech-to-text systems, including dependency on internet-based cloud services, privacy concerns,

high operational costs, and poor handling of musical or multilingual content. Through the implementation of a Zero-Cloud Architecture, all audio processing is performed locally on the user's device, ensuring enhanced privacy, security, and offline accessibility. This makes the system particularly suitable for students, researchers, independent creators, and users handling sensitive audio recordings.

One of the major innovations of the project is the introduction of the Lyrics Mode, which uses Voice Activity Detection and Temporal Gap Analysis to intelligently organize lyrical content into structured lines and verses. Unlike conventional ASR systems that generate continuous blocks of text, the proposed solution improves readability and significantly reduces manual formatting effort for songs, poetry, and rhythmic speech content.

## References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.