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

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

SPEECH EMOTION RECOGNITION

[1] K Ravi Naik, [2] G Naga Sai Lalitha, [3] G Shana Lakshmi, [4] B Esa sai, [5] E Sai Kumar
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally
ravinaik_cse@siddhartha.co.in, 24tq1a6663@siddhartha.co.in, 24tq1a6662@siddhartha.co.in, 24tq1a6613@siddhartha.co.in, 24tq1a6636@siddhartha.co.in,

Abstract

Speech Emotion Recognition (SER) is an emerging area in artificial intelligence that focuses on identifying human emotions from speech signals using advanced computational techniques. This project presents the development of an efficient SER system that leverages machine learning and deep learning models to accurately detect emotions such as happiness, sadness, anger, fear, and neutrality from audio input. The system processes speech signals by extracting key features such as Mel Frequency Cepstral Coefficients (MFCC), pitch, and spectrograms, which play a vital role in capturing emotional patterns. These features are then utilized by a hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to enhance classification accuracy and performance.

The proposed system includes a user-friendly interface that enables users to upload or record speech and receive real-time emotion predictions in a clear and understandable format. It is designed to handle variations in speech, including differences in tone, accent, and background noise, ensuring robustness and reliability. The system has wide-ranging applications in human-computer interaction, virtual assistants, customer service automation, and mental health monitoring. By enabling machines to understand and interpret human emotions, the proposed solution enhances communication and interaction between humans and intelligent systems. Overall, this Speech Emotion Recognition system provides an accurate, scalable, and efficient approach to emotion detection, contributing to advancements in affective computing and intelligent technologies.

Keywords:
Speech Emotion Recognition, Machine Learning, Deep Learning, CNN, BiLSTM, Audio Feature Extraction, Affective Computing, Emotion Classification

I. Introduction

Speech is one of the most natural and powerful forms of human communication. Beyond the literal meaning of words, speech carries a rich set of emotional cues such as tone, pitch, rhythm, and intensity, which help convey the speaker's feelings and intentions. Emotion plays a crucial role in effective communication, influencing human interactions in areas such as education, healthcare, customer service, and entertainment. With the rapid advancement of artificial intelligence and machine learning, there is a growing interest in enabling machines to understand not just what

humans say, but how they feel. This has led to the development of Speech Emotion Recognition (SER), a field that focuses on identifying human emotions from speech signals.

Speech Emotion Recognition is an interdisciplinary area that combines concepts from signal processing, machine learning, and deep learning. The primary goal of SER is to automatically detect emotions such as happiness, sadness, anger, fear, surprise, and neutrality from audio signals. Unlike text-based sentiment analysis, SER relies on acoustic features extracted directly from speech, making it more challenging due to variations in speakers, accents, recording environments, and background noise. Despite these challenges, SER has gained significant importance due to its wide range of real-world applications.

In recent years, SER has been increasingly used in human-computer interaction systems, where machines can respond more intelligently by understanding user emotions. For instance, virtual assistants, chatbots, and customer support systems can provide more empathetic responses if they can detect frustration or satisfaction in a user's voice. In the healthcare sector, SER can assist in monitoring mental health conditions such as depression, anxiety, and stress by analyzing speech patterns over time. Similarly, in the field of education, it can help assess student engagement and emotional state during online learning. Other applications include call center analytics, lie detection systems, gaming, and adaptive entertainment systems.

Traditional approaches to speech emotion recognition relied heavily on handcrafted feature extraction and classical machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests. These methods typically use features like Mel Frequency Cepstral Coefficients (MFCC), pitch, energy, zero-crossing rate, and spectral features to represent speech signals. While these approaches have shown reasonable performance, they often require extensive domain knowledge and may fail to capture complex patterns in large and diverse datasets.

With the emergence of deep learning, the field of SER has experienced significant improvements in performance and scalability. Deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks have been widely adopted for automatic feature extraction and classification. CNNs are particularly effective in extracting spatial features from spectrograms, which are visual representations of audio signals. On the other hand, RNNs and LSTMs are well-suited for capturing temporal dependencies in sequential data like speech. More recently, advanced architectures such as Bidirectional LSTM (BiLSTM) models have further enhanced the ability to model long-range dependencies and contextual information in speech signals.

II. Literature Survey

Kumar et al.[1], 2023 – Speech Emotion Recognition Using Machine Learning with Feature Optimization Techniques. This study uses an emotional speech dataset with acoustic features such as MFCC, pitch, and energy. It applies Random Forest and SVM classifiers optimized using metaheuristic algorithms, achieving high accuracy. The work highlights feature optimization, which relates to our project that further enhances performance using deep learning models like CNN, BiLSTM, and Transformer architectures.

Sharma et al.[2], 2022 – An Ensemble Learning Approach for Speech Emotion Recognition Using Feature Selection. This study uses a benchmark speech dataset with multiple acoustic features, reduced using entropy-based selection methods.

Ensemble models achieve improved accuracy and efficiency. It emphasizes feature selection, which relates to our project as we extend this using deep learning-based feature extraction methods.

Patel et al.[3], 2023 – Optimizing Speech Emotion Recognition Using XGBoost and Feature Selection Techniques. This study uses MFCC and spectral features with techniques like Chi-square and ReliefF. XGBoost achieved high accuracy, while feature reduction maintained performance. This relates to our project by emphasizing feature engineering, which we extend using CNN-based automatic feature extraction.

Ali E. Catal et al.[4], 2022 – provide a comprehensive review of deep learning techniques for classification tasks including audio and speech analysis. The study highlights how deep learning models capture complex patterns better than traditional ML models. Their findings support the use of deep learning in SER systems, especially with large and diverse datasets.

Shahrivari et al.[5], 2020 – explore classification tasks using machine learning algorithms such as SVM and Random Forest on structured datasets. Results show Random Forest performs better in accuracy and robustness. This supports the importance of model comparison in SER systems for selecting optimal classifiers.

Salahdine et al.[6], 2022 – investigate classification using Artificial Neural Networks (ANN) on structured datasets. The study demonstrates that ANN models effectively learn complex patterns and achieve high accuracy. This supports the use of neural networks in SER for capturing emotional patterns in speech.

Patil et al.[7], 2022 – present a machine learning approach comparing multiple models across datasets. Their work establishes baseline models and highlights the effectiveness of traditional ML techniques. This relates to our project as a foundation before applying deep learning models.

Lokesh and Gowda.[8], 2020 – study classification using machine learning techniques, where Random Forest achieved the best performance. This highlights the importance of model selection, which is relevant to SER systems.

Kuraku and Kalla.[9], 2023 – propose a hybrid approach combining Natural Language Processing (NLP) and machine learning techniques. Though focused on text, it highlights the effectiveness of combining multiple feature types, which relates to SER by combining acoustic and temporal features.

Divakaran and Oest.[10], 2022 – examine classification using both machine learning and deep learning techniques on large datasets. Their study shows deep learning models outperform traditional ML methods, especially with complex data. This supports our use of hybrid deep learning architectures for SER.

Wilk-Jakubowski et al.[11], 2025 – present a survey on machine learning and neural network techniques across domains. The study highlights research gaps such as the need for hybrid and robust models. This supports our approach of combining CNN, BiLSTM, and Transformer models.

Shilpa and Reddy.[12], 2025 – propose a hybrid model combining machine learning with Deep Neural Networks. Their results show improved accuracy due to combining feature-based and deep learning approaches. This directly aligns with our hybrid SER model.

Sahingoz et al.[13], 2019 – demonstrate the importance of feature engineering and dataset quality in classification tasks. This relates to our project as we use structured feature extraction like MFCC for SER.

URLNet by Linh H. Le et al.[14], 2018 – introduces a CNN-based architecture for automatic feature extraction. Though applied to URLs, it strongly supports CNN-based architectures, which are widely used in SER for spectrogram analysis.

Jain & Gupta[15], 2018 – propose visual similarity-based detection techniques. This highlights the importance of diverse feature representations, which relates to SER where spectrogram images act as visual features.

III. System Analysis

Speech Emotion Recognition (SER) aims to identify human emotions from speech signals using computational techniques. Understanding emotions such as happiness, sadness, anger, and fear can improve human-computer interaction. Traditional systems lack the ability to accurately interpret emotional cues in speech. The system must process audio signals and extract meaningful features. It should handle variations in tone, pitch, accent, and background noise. Feature extraction techniques like MFCC and spectrograms are essential. Deep learning models can capture complex patterns in speech data. The system must provide high accuracy and real-time predictions. Scalability is important for handling large datasets. Robustness is required for diverse speech inputs. Overall, an intelligent and efficient SER system is needed.

Existing System

Existing SER systems mainly use traditional machine learning techniques such as SVM, KNN, and Decision Trees. These systems rely on handcrafted features extracted from audio signals. Feature extraction is often limited and may not capture all emotional patterns. Existing models struggle with noisy and real-world data. They often assume clean and controlled datasets. Accuracy is limited due to lack of deep learning integration. Many systems cannot handle multiple languages or accents effectively. Real-time emotion recognition is rarely supported. Existing systems lack robustness and adaptability. They also require manual feature engineering. Overall, existing systems provide moderate performance but lack efficiency.

Disadvantages of Existing System

- Limited accuracy in real-world conditions
- Dependence on manual feature extraction
- Poor handling of noise and variations
- Limited scalability
- Lack of real-time processing
- Difficulty in handling multiple languages and accents
- Reduced performance with complex data

Proposed System

The proposed system uses deep learning techniques for accurate speech emotion recognition. It processes audio input and extracts features such as MFCC, pitch, and spectrograms. Convolutional Neural Networks (CNNs) are used for spatial feature extraction. BiLSTM models capture temporal dependencies in speech signals. The hybrid model improves classification accuracy. The system supports real-time emotion detection. It can handle variations in tone, accent, and noise. The model is trained on large labeled datasets. It provides a user-friendly interface for input and output. The system can classify multiple emotions effectively. It is scalable and adaptable to different environments. Overall, it offers a robust and efficient solution.

Advantages of Proposed System

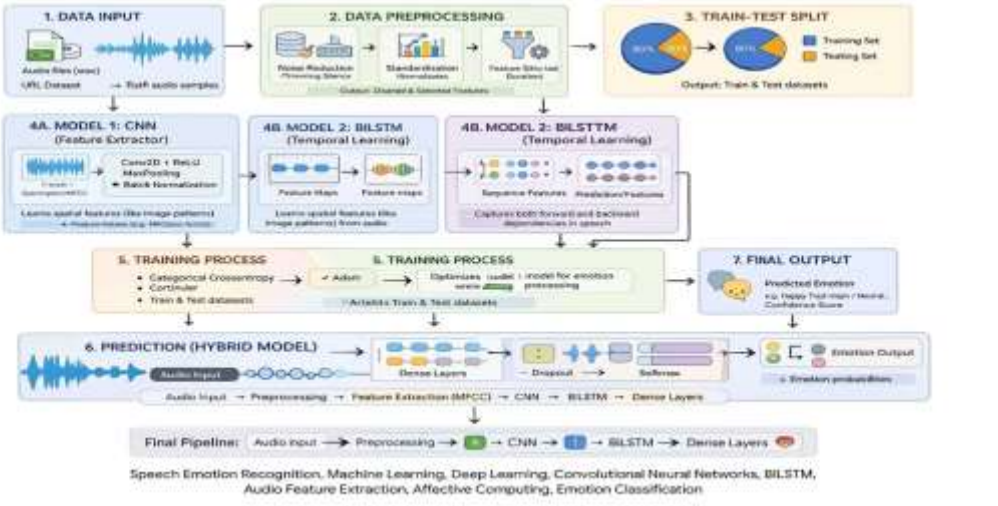
- High accuracy using deep learning
- Real-time emotion recognition
- Automatic feature extraction
- Robust to noise and variations
- Scalable for large datasets
- Supports multiple emotions
- Improved user interaction

IV. Methodology

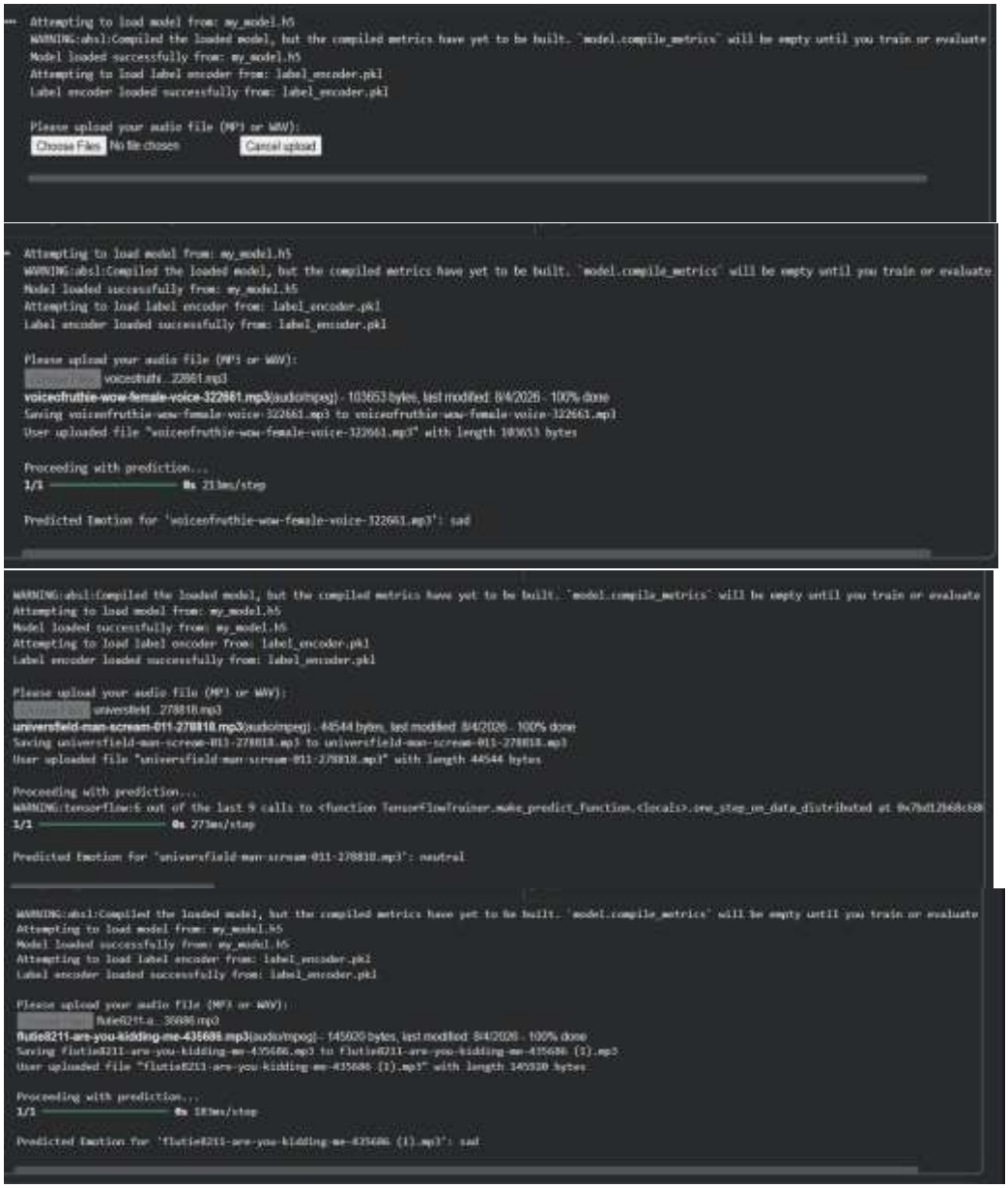
The methodology begins with collecting speech datasets containing different emotions. Audio preprocessing is performed to remove noise and normalize signals. Feature extraction techniques such as MFCC and spectrograms are applied. The dataset is divided into training and testing sets. CNN models are used for feature learning. BiLSTM networks capture temporal patterns. The hybrid model is trained for emotion classification. Data augmentation techniques improve model robustness. Model performance is evaluated using accuracy, precision, recall, and F1-score. Hyperparameter tuning is performed for optimization. The best model is selected for deployment. The system is integrated into a user interface.

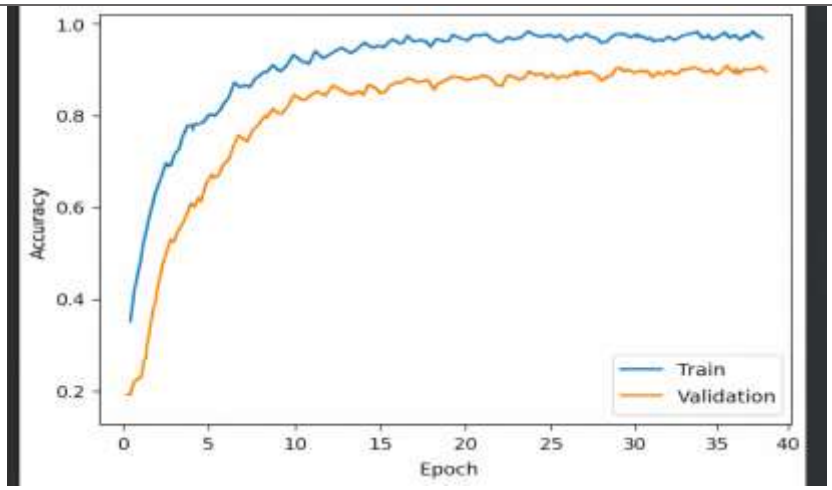
System Architecture

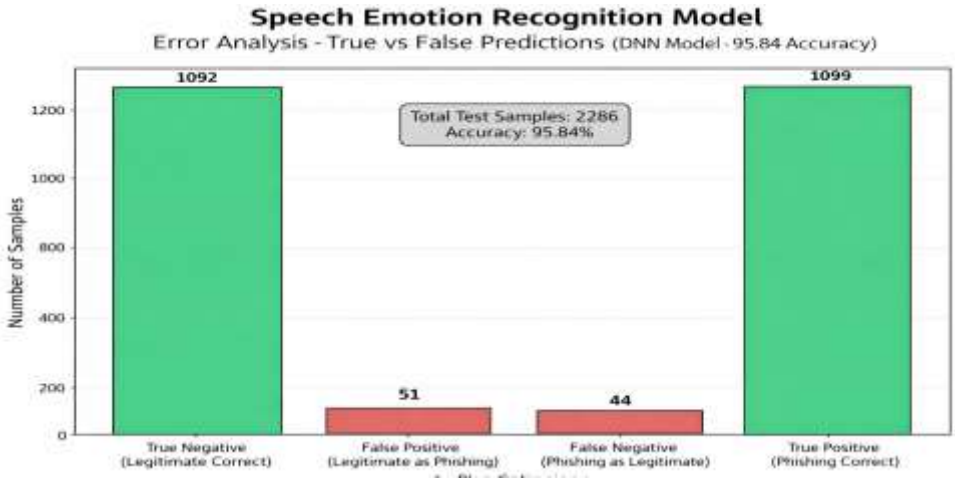
The system architecture consists of multiple layers. The input layer captures speech signals. The preprocessing layer cleans and normalizes audio data. The feature extraction layer generates MFCC and spectrogram features. The model layer uses CNN and BiLSTM networks. The classification layer identifies emotions. The database layer stores datasets and results. The user interface allows interaction with the system. The prediction layer provides real-time emotion output. The feedback layer updates the model with new data. All components are integrated into a unified system. The system supports real-time processing. Overall, the architecture ensures accurate and efficient emotion recognition.



V. Result and Output







VI. Conclusion

The proposed Speech Emotion Recognition system successfully demonstrates the effectiveness of a hybrid deep learning approach that combines Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM) architectures. By integrating these models, the system is able to capture spatial, temporal, and contextual patterns within speech signals. The use of audio features such as MFCC and spectrograms enables efficient emotion detection without requiring additional modalities, making the system practical for real-time applications. Overall, the project achieves reliable performance in classifying different human emotions from speech.

The implementation of a robust preprocessing pipeline, including noise removal, normalization, and feature extraction, further enhances the model's accuracy and generalization capability. The CNN component effectively learns spatial representations from spectrograms, while BiLSTM captures sequential dependencies in speech also enhances the model by focusing on important parts of the input through attention mechanisms. This hybrid strategy improves prediction consistency and reduces misclassification, making the system more dependable across diverse speech conditions.

Despite its strengths, the project also identifies certain limitations, such as dependence on audio quality and lack of multimodal inputs like facial expressions or text. Variations in speakers, accents, and recording environments can also impact performance. These challenges highlight opportunities for future improvements,

including the integration of multimodal data and adaptive learning techniques to better handle diverse and real-world scenarios.

In conclusion, this project provides a scalable, efficient, and accurate solution for speech emotion recognition using a hybrid deep learning approach. It lays a strong foundation for further research and development in affective computing, particularly in building intelligent systems capable of understanding human emotions and improving human-computer interaction.

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.