

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**

Research Paper

REDUCING DATA VOLUME IN NEWS TOPIC CLASSIFICATION USING DEEP LEARNING

¹A Praveen, ²P Saikumar, ³Ch Shravan Kumar, ⁴N Nikhil reddy ⁵KV Shiva Sai Ganesh

¹AssistantProfessor, ²³⁴⁵Students

Department of CSE(AI & ML)

Siddhartha Institute of Technology & Sciences, Narapally

aiml.hod@siddhartha.co.in, 23TQ1A6627@siddhartha.co.in, 23TQ1A6646@siddhartha.co.in,

23TQ1A6603@siddhartha.co.in, 23TQ1A6618@siddhartha.co.in.

Abstract

The exponential growth of digital news content has posed significant challenges in efficiently processing and classifying large volumes of textual data. News topic classification is essential for organizing, filtering, and retrieving relevant information from extensive news datasets. However, high-dimensional text data increases computational complexity, memory usage, and training time in deep learning models. This project focuses on reducing data volume in news topic classification while preserving classification accuracy through advanced deep learning techniques.

The proposed approach incorporates effective text preprocessing and dimensionality reduction methods, including tokenization, stop-word removal, stemming, and feature selection, to eliminate redundant and irrelevant information from news articles. By optimizing the input data size, the system enhances computational efficiency and reduces resource consumption. A deep learning model is then utilized to classify news articles into predefined categories such as politics, sports, technology, and business.

Experimental results indicate that data volume reduction significantly improves training speed and overall efficiency with minimal impact on classification performance. The study emphasizes the importance of efficient data preprocessing and feature optimization in large-scale text classification tasks. The proposed system can be effectively applied in news recommendation systems, automated journalism analysis, and large-scale information retrieval applications.

I. Introduction

In recent years, the rapid expansion of the internet and digital communication technologies has led to an exponential increase in the generation and distribution of news content. Thousands of news articles are published every minute across various online platforms such as news websites, blogs, and social media networks. This continuous surge of information has created a strong need for intelligent systems that can automatically organize and categorize news data efficiently. As a result, news topic classification has become a significant research area in Natural Language Processing (NLP) and text mining.

News topic classification involves automatically assigning news articles to predefined categories such as politics, sports, business, technology, entertainment, and health

based on their content. This process plays a vital role in improving information accessibility and supports applications like news aggregation platforms, recommendation systems, content filtering, and digital libraries. By automating the classification process, these systems reduce manual effort and enhance user experience.

Traditional text classification techniques rely on machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Decision Trees. These approaches typically use feature extraction methods like Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) to convert textual data into numerical representations.

II. Literature Survey

Several researchers have explored news topic classification and data optimization techniques using machine learning and deep learning approaches.

Serreli et al. (2024) proposed a deep learning framework focused on reducing data volume in news topic classification. Their approach applied preprocessing and feature reduction techniques to eliminate redundant textual information. The results showed improved computational efficiency while maintaining high classification accuracy, highlighting the importance of data reduction in large-scale systems .

Daud et al. (2023) utilized optimized machine learning models for classifying online news articles. They employed feature extraction methods such as TF-IDF along with algorithms like Support Vector Machines and Random Forest. Their study demonstrated improved classification accuracy and performance, emphasizing the effectiveness of optimized machine learning techniques in news categorization .

Sunagar et al. (2021) explored traditional machine learning algorithms including Naïve Bayes, Decision Trees, and Support Vector Machines for news classification. The study showed that these methods can effectively categorize news articles when combined with appropriate preprocessing and feature selection techniques .

Kuzman and Ljubešić (2025) introduced a teacher–student framework using large language models (LLMs) for text classification without requiring manually annotated data. The teacher model generated pseudo-labels to train a smaller student model, reducing dependency on labeled datasets while maintaining competitive performance

Rodrigues and Chiplunkar (2022) developed a big data framework for topic classification and sentiment analysis using Twitter data. Their system used distributed processing techniques to efficiently handle large-scale textual datasets, emphasizing scalability in text classification systems .

Iman et al. (2017) conducted a longitudinal study on topic classification in Twitter data, focusing on how topics evolve over time. Their findings highlighted the need for adaptive classification models capable of handling dynamic and continuously changing datasets

Lee et al. (2011) proposed a real-time system for classifying trending topics on Twitter. The study emphasized challenges in handling streaming data and the need for efficient and fast classification methods for large-scale text processing

III. System Analysis

The rapid growth of digital news content has created a need for efficient systems to manage and classify large volumes of textual data. News topic classification helps in organizing and retrieving relevant information automatically. However, handling high-dimensional text data increases computational complexity and processing time. Traditional systems struggle with scalability when dealing with large datasets. Deep learning models provide better accuracy but require high computational resources. Redundant and irrelevant data further increases inefficiency. Therefore, there is a need for a system that reduces data volume while maintaining classification performance. The proposed system focuses on optimizing data through preprocessing and feature selection. It aims to improve efficiency, scalability, and accuracy. This system is suitable for real-time applications and large-scale news platforms.

Existing System

The existing systems for news topic classification mainly rely on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines, and Decision Trees. These systems use feature extraction techniques like Bag-of-Words (BoW) and TF-IDF to convert text into numerical form. Although these methods provide reasonable accuracy, they often generate high-dimensional feature spaces. This leads to increased computational cost and memory usage. Many existing systems process the entire dataset without removing redundant or irrelevant information. As a result, training time becomes longer and model efficiency decreases. Additionally, these systems require manual feature engineering. They also struggle with large-scale and real-time data processing. Overall, existing systems lack efficiency and scalability.

Disadvantages of Existing System

- High computational complexity due to large feature space
- Increased memory usage and storage requirements
- Longer training and processing time
- Presence of redundant and irrelevant data
- Requires manual feature engineering
- Poor scalability for large datasets

Proposed System

The proposed system introduces an efficient approach for reducing data volume in news topic classification using deep learning techniques. It incorporates text preprocessing methods such as tokenization, stop-word removal, and stemming to clean the data. Feature selection and dimensionality reduction techniques are applied to retain only relevant information. This reduces the size of the dataset significantly. A deep learning model is then used to classify news articles into predefined categories. The system improves computational efficiency and reduces training time. It also

enhances model performance by focusing on meaningful features. The proposed system is scalable and suitable for large datasets.

Advantages of Proposed System

- Reduced data volume and improved efficiency
- Lower computational cost and memory usage
- Faster training and processing time
- Improved classification accuracy
- Eliminates redundant and irrelevant data
- Better scalability for large datasets

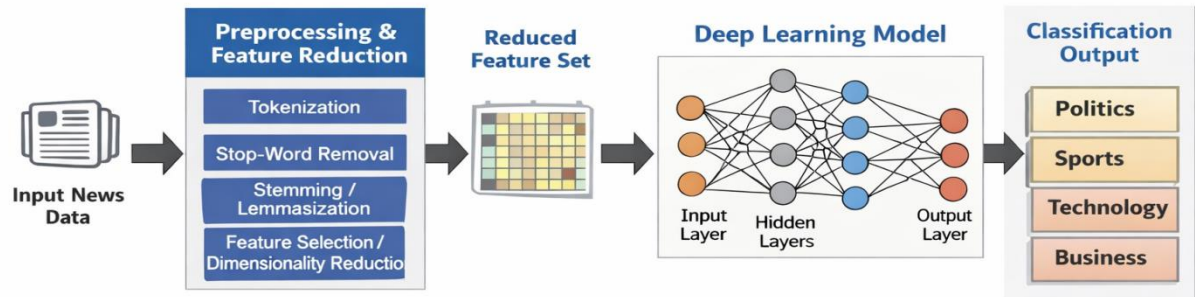
IV. Methodology

The methodology begins with collecting a dataset of news articles from various sources. The data is then preprocessed using techniques such as tokenization, stop-word removal, and stemming to clean and standardize the text. Next, feature extraction methods are applied to convert text into numerical representations. Dimensionality reduction and feature selection techniques are used to reduce data size by removing irrelevant features. The processed data is then fed into a deep learning model such as CNN or LSTM for classification. The model is trained using labeled data to learn patterns and relationships. After training, the model is tested on unseen data to evaluate performance. Metrics such as accuracy and precision are used for evaluation. Finally, the system outputs the classified news category efficiently. The overall process ensures improved performance with reduced data volume.

System Architecture

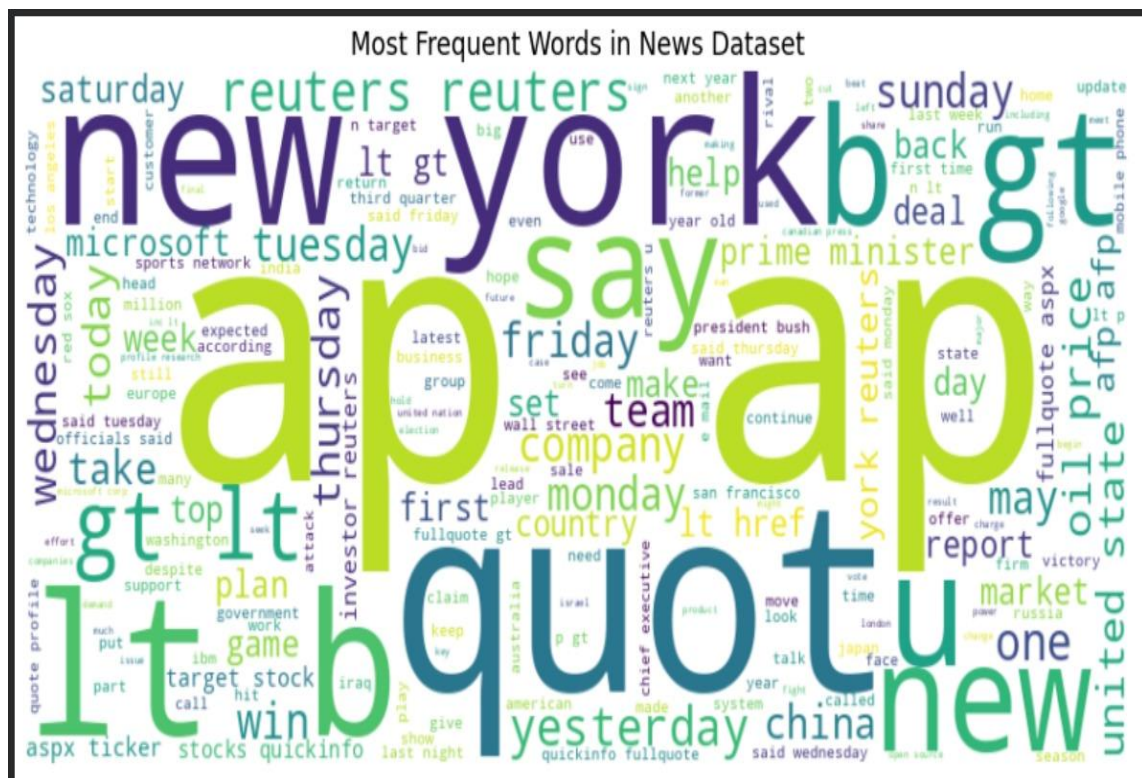
The system architecture for *Reducing Data Volume in News Topic Classification using Deep Learning* consists of multiple stages designed to efficiently process and classify large-scale textual data. Initially, news articles are collected from various sources such as online news websites and datasets. The raw text data is then passed through a preprocessing module where tokenization, stop-word removal, stemming, and normalization are performed to clean and standardize the data. After preprocessing, feature extraction techniques such as TF-IDF or word embeddings are applied to convert text into numerical representations.

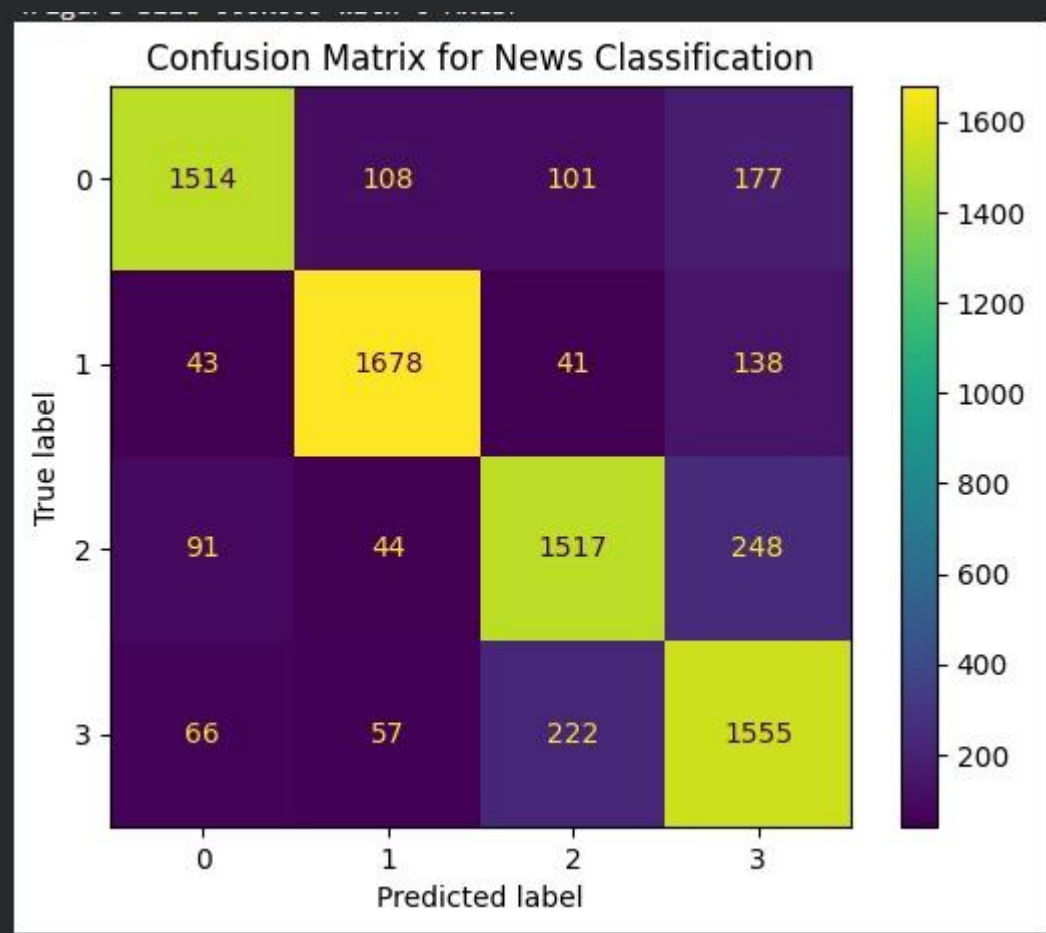
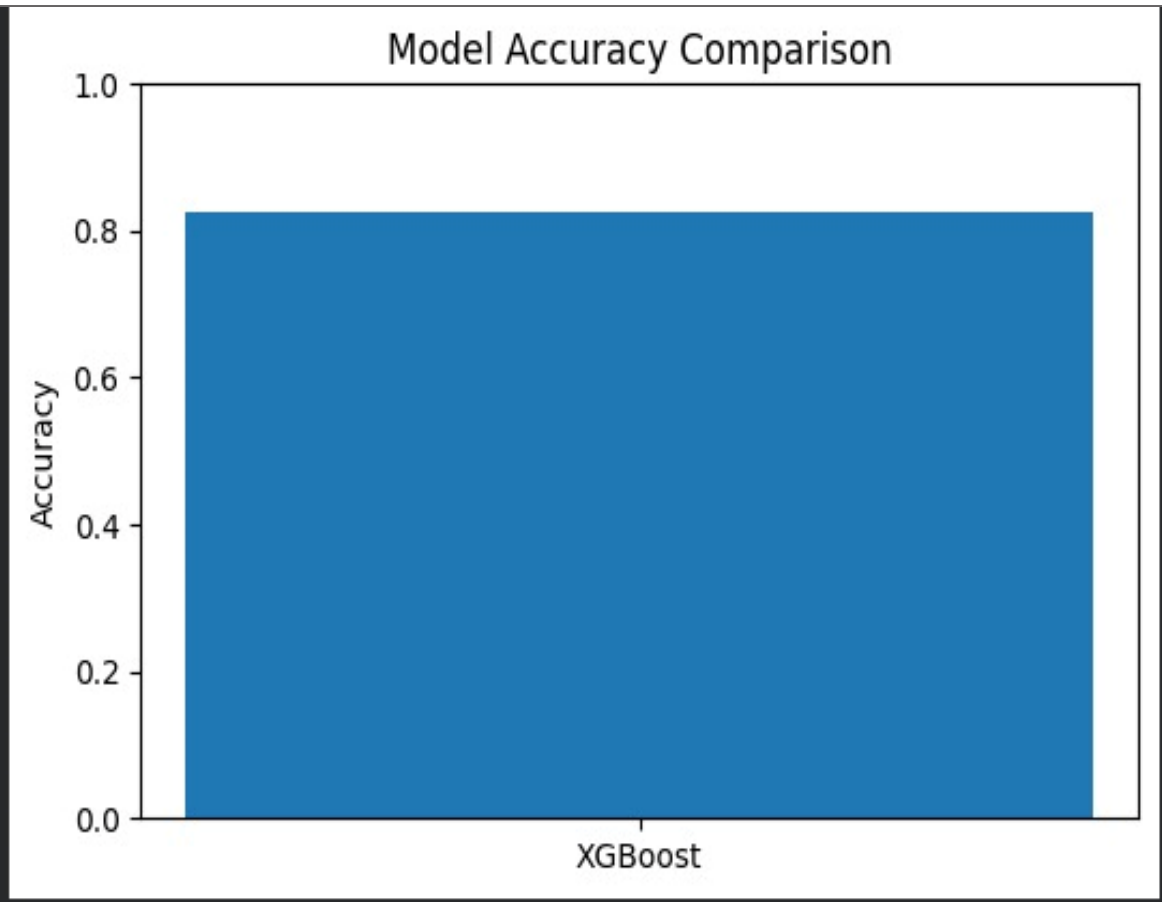
Reducing Data Volume in News Topic Classification Using Deep Learning

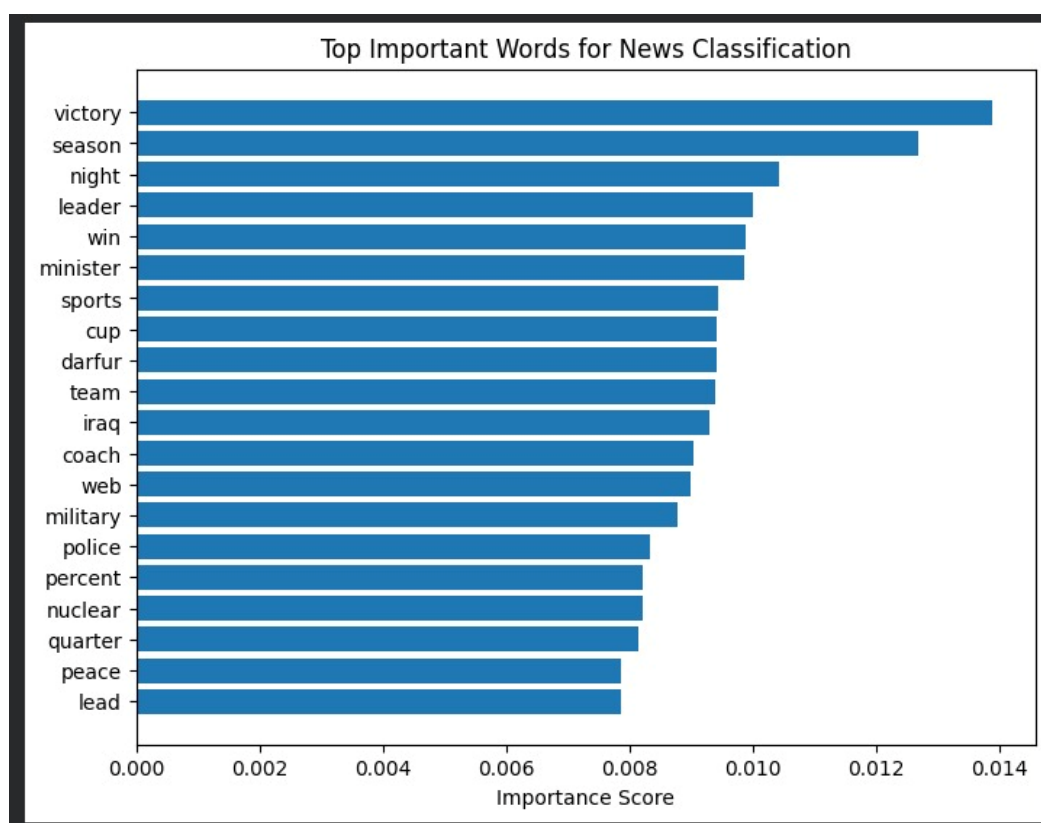


Model Architecture

V. Result and Output







VI. Conclusion

In this study, a deep learning-based framework for reducing data volume in news topic classification was successfully proposed. The approach focuses on minimizing the size of textual datasets while preserving essential information required for accurate classification. By applying effective text preprocessing techniques such as tokenization, stop-word removal, stemming, and feature selection, redundant and irrelevant information was eliminated, resulting in an optimized and reduced feature set.

The processed dataset was then used to train a deep learning model capable of learning meaningful patterns and accurately classifying news articles into predefined categories. The results demonstrate that reducing data volume significantly improves computational efficiency by lowering training time, memory usage, and overall model complexity. Importantly, this reduction does not lead to a major loss in classification accuracy, proving that efficient feature selection can maintain performance while enhancing system efficiency.

References

- [1] Kumar, R. D., Prudhvraj, 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.