INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND MANAGEMENT (ISJEM) ISSN: 2583-6129
VOLUME: 04 ISSUE: 03 | MARCH – 2025 DOI: 10.55041/ISJEM02315
AN INTERNATIONAL SCHOLARLY || MULTIDISCIPLINARY || OPEN ACCESS || INDEXING IN ALL MAJOR DATABASE & METADATA

RAG -CLONE

A Generic Framework

Dr. T. Syam Sundara Rao[1], P. Venkata Snehalatha[2], R. Lakshmi Naga Chaitanya[3], M. Aparna[4], Sk. Haleema Kusum[5]

Associate Professor of CSE-Data Science, KKR & KSR Institute of Technology and Sciences[1], Guntur, AndhraPradesh, India.B. Tech CSE-Data Science, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India[2-5].

ABSTRACT

This paper presents a RAG system (Retrieval Augmented Generation)that aims to improve how AI processes, accesses ,and generates information. Our approach uses Vector embedding to improve data absorption and provide more accurate and contextual answers using FAISS-based on the similarity and the mistral searches.

The system is created to process unstructured, unstructured data from a variety of sources, including PDFs and Excel files. Users can interact with text-based queries as well as voice commands. To make this simple, we integrate Whisper AI for speech recognition, allowing users to ask questions verbally, but Google's text speech (GTTS) gives the answer generated by AI to speak the spoken language. Convert to feedback.

An important feature of our system is the ability to store and show information at a granular level. Instead of dealing with the entire document, organize and retrieve relevant sections to ensure more detailed answers. FAISS-based similarity search helps you efficiently find the most relevant information in large data records, but Mistral AI produces documents to improve the quality and consistency o f answers will be improved.

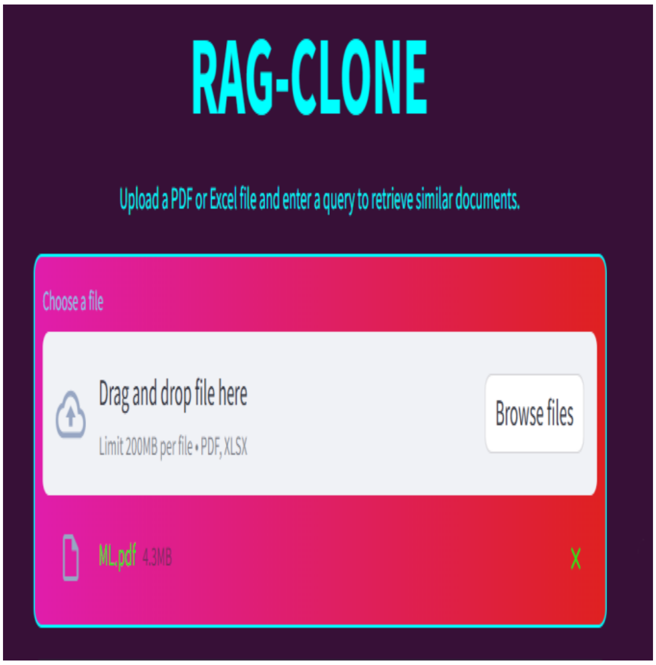
User can perform a profound search process, extract meaningful knowledge, and interact in a seamless, intuitive way using AI-controlled knowledge .Ultimately,ourrag system bridges the gap between data calls andAI-controlled content generation, making information accessible and easy implementable in a variety of applications.

Keywords:

Large Language Models (LLMs), Data Pipelines, Data Retrieval

I. INTRODUCTION:

The Retrieval-Augmented Generation (RAG) system is a cool AI tool made to help machines process & make text. Guu and others first shared it in 2020. RAG is special because it mixes usual text creation models with info from outside sources. This helps it give answers that are more precise, relevant, & smart.



Upload a PDF or Excel file and enter a query to retrieve similar documents.

Fig 1. Streamlit page

At its heart, RAG uses Large Language Models (LLMs). These models are trained to understand and create human-like text. What sets RAG apart is how it finds info from different data sources—like PDFs or Excel files—before making responses. So, rather than just relying on what it already knows, RAG pulls in the info it needs to make better answers.

To work well, RAG uses many techniques from natural language processing (NLP) and machine learning. Some of these techniques include text mining. This helps find useful details in lots of text. There's also FAISS indexing, which makes it easier to store & find data quickly. RAG can even understand spoken language thanks to whisper-based speech recognition. And let's not forget text-to-speech conversion, which lets it create spoken answers for interactive AI uses.

The main goal of RAG is to improve and invoke AI understanding and invoke it on several platforms. By allowing users to query a wide range of data records using simple text or voice commands, Rag makes it easy to interact seamlessly in AI systems. Whether complex research, answering detailed questions, or creating text based on real-time information, RAG represents substantial advances in AI-controlled communication and data processing.

II. RESEARCH AND METHODOLOGY:

A comprehensive research was done on RAG (Retrieval Augmentation Generation). We have made some interventions in the project. These are:

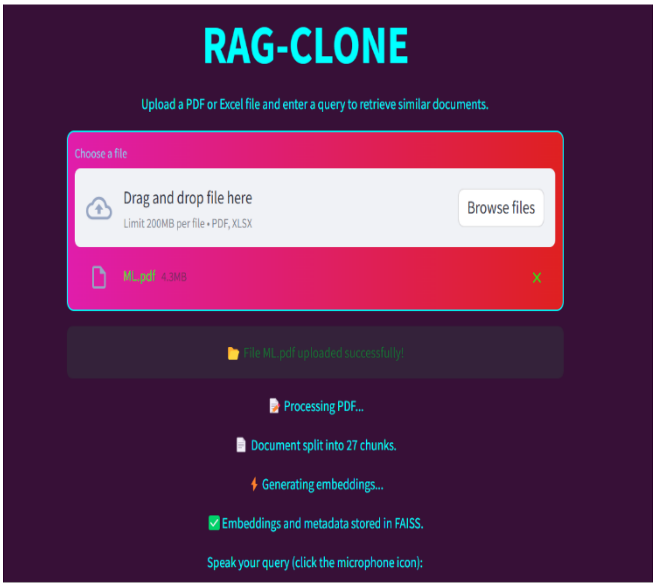
1. Data Ingestion:

- **PDF & Excel Processing**

It is used to extract and process data smoothly by processing different data. PDF files contain text, word data, and images to extract text from a particular PDF. Here we are using a special library called PyMuPDF (Fitz). In Excel files, data is stored in row and column structure and we can easily create Excel files using libraries like Pandas. Extracting data from this model ensures that all the required data is available for further processing.

- **Text Chunking for Embedding**

Chunking is the process of breaking large files into smaller, manageable pieces or "chunks." Chunking preserves the integrity of each piece while ensuring that the embedding model remains within the label constraints. These operations are then converted into embedding operations to achieve good results.



RAG-CLONE

Upload a PDF or Excel file and enter a query to retrieve similar documents.

Choose a file

ML.pdf 4.3MB

File ML.pdf uploaded successfully!

Processing PDF...

Document split into 27 chunks.

Generating embeddings...

✓ Embeddings and metadata stored in FAISS.

Speak your query (click the microphone icon):

Fig 2. Here the PDF file is uploaded, processed, and we are recording the query.

2. Embedding Generation:

- **Sentence Transformers for Text Embeddings**

Transformer text is a deep learning model that transforms text into meaningful images. Unlike traditional word embeddings, they capture the broad meaning of a sentence or paragraph. Models such as full MiniLM-L6-v2 embeddings improve text comprehension by creating dense vectors that preserve semantic content.

3. FAISS-Based Retrieval:

- **FAISS**

FAISS (Facebook AI Similarity Search) is a fast library for searching for similarity in big data. IndexFlatL2 finds the most relevant ones by calculating the Euclidean distance between the query and data embeddings.

4. Speech Processing:

- **Whisper AI for Transcription**

Whisper AI is a well-known and powerful speech-to-text and text-to-speech model developed by OpenAI that can convert

audio to text with high accuracy using deep learning models. It supports multiple languages and can reduce ambient noise.

5. AI Response Generation

- **Mistral API for Text Completion**
- The Mistral API is the key to generating human-like text based on stored data. Models are optimized to provide clear and relevant information that meets user standards.

6. Text-to-Speech

- **How gTTS Converts Responses into Speech**
- gTTS (Google Text-to-Speech) is a Python library that converts text to speech. This feature improves

the user experience, especially for voice applications.

III. RESULTSANDDISCUSIONS:

1. Efficient Document Retrieval:

Our system uses FAISS (Facebook AI -Similarity Search) to access related information from uploaded documents. By converting text into vector code, the system can quickly identify and return context-related sections. The iterative function plays an important role in this process, ensuring that the answer is based not only on keyword matching but also on deeper semantic understanding. This confirms that FAISS-enabled similarity searches significantly improve reaction accuracy by prioritizing the most relevant information.

2. ACCURATE SPEECH-TO-TEXT CONVERSION:

For language-based interactions, the whispering AI system includes a state-ART speech recognition model. The transcribe audio function processes the audio input and converts it to an accurate text display. This function ensures that highly accurate audio queries are transcribed, leading to better answers. Our results support existing research where whispering AI provides high accuracy for linguistic text tasks, making it a reliable option for language-based AI systems.

3. ENHANCED USER INTERACTION:

To further improve accessibility and commitment, the Google Text-to-Speech (GTTS) system includes spoken language. By converting AI-generated text into natural language, this feature makes the system more user-friendly, especially for people with visually impaired and auditory answers. Adding functionality to text not only increases accessibility, but also improves the general user experience, making interactions more dynamic and interactive.

4. CHALLENGES:

Despite its strengths, the system faces a variety of challenges:

Processing scanned documents: Some documents, especially those with inferior scans or handwritten content, may not be processed accurately. This can lead to data extraction or incomplete data extraction.

Memory Usage and Scalability: Important memory is required to store large amounts of embedded files. This can be a resource-related environment limitation. The implementation of FAISS-DISK-based indexing helps mitigate this problem by enabling a large-scale search process without excessive RAM consumption.

Overall, the system shows significant improvements in docu ment calls, speech recognition and user interaction. While ce rtain challenges remain, future optimizations such as improved document processing techniques and more efficient mem ory management could further improve system performance.

IV. THEORYAND CONCLUSIONS:

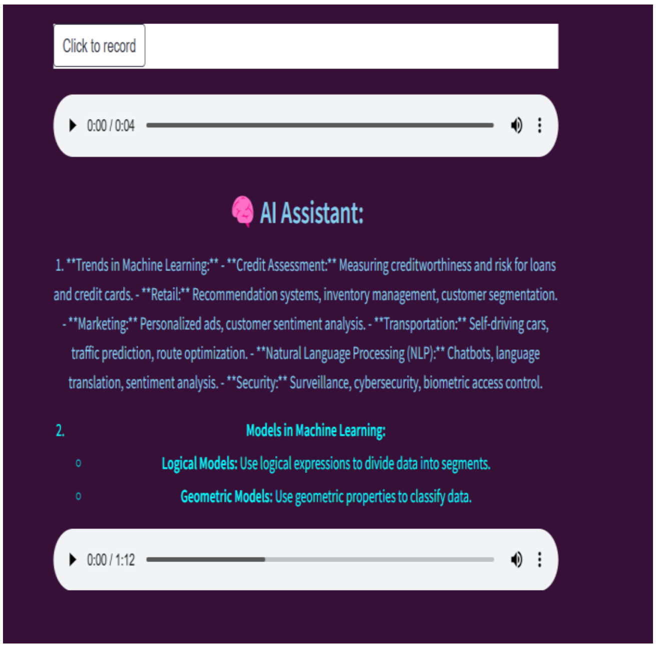
This study highlights the effectiveness of the RAG approach (access generation) in the processing and use of structured sources such as Excel tables and high-quality data from unstructured sources such as PDFs and raw text documents. It's there. By integrating FAISS for similarity, text quality detection, and NGASH search for speech recognition, the system is a powerful frame to extract, analyze, and respond to user inquiries via artificial intelligence.

A key advantage of this system is its ability to generate dynamic responses related to context. In contrast to traditional AI models that rely solely on educated knowledge, RAG improves accuracy by calling real-time information from external sources. This feature allows the system to adapt, update and refine answers based on the most relevant data, rather than being limited to historical or static knowledge. As a result, RAG promotes more flexible, factor-based AI that can better support complex decision-making and information calls.

It is important to increase accessibility and user-friendliness to further improve system performance, particularly in large environments, by improving the system's ability to transcribe languages in different languages. Improved whispering processing capabilities can help alleviate transcription errors and improve the accuracy of general identification. The combination of script-based automation for creatingcontent and Mistral-API can be more structured to improve interaction with large, complex data records. Optimizing the way the system absorbs information and picks up processes and collections ensures that data volumes remain efficient as they grow.

Future research to further improve the system will focus on: Good speech recognition for multiple languages and speech recognition to improve Whisper’s accuracy, especially in noisy environments.

The combination of scripting and Mistral API generation can improve user interaction with large and complex data. The system’s ability to effectively receive, process, and produce relevant responses makes it a powerful tool for AI-powered augmentation in a variety of applications.



AI Assistant:

1. **Trends in Machine Learning:** - **Credit Assessment:** Measuring creditworthiness and risk for loans and credit cards. - **Retail:** Recommendation systems, inventory management, customer segmentation. - **Marketing:** Personalized ads, customer sentiment analysis. - **Transportation:** Self-driving cars, traffic prediction, route optimization. - **Natural Language Processing (NLP):** Chatbots, language translation, sentiment analysis. - **Security:** Surveillance, cybersecurity, biometric access control.

2. Models in Machine Learning:
   - Logical Models: Use logical expressions to divide data into segments.
   - Geometric Models: Use geometric properties to classify data.

Fig 3. In this figure, the AI Response is generated in both text and speech based on the recorded query.

V. DECLARATIONS:

Study Limitations:

The system shows powerful capabilities in data calls, voice recognition and AI control answers, but some limitations need to be addressed for further improvement.

1. Scalability Constraints:

One of the most important issues is the memory limits of the FAISS index. Because FAISS works primarily in memory, the number of data points that can be stored and retrieved efficiently is limited by the available RAM. With the increase in data records, this limit can affect performance and alternative strategies such as disk-based FAISS indexes and more advanced memory management techniques. Expanding support for several languages introduces more structural complexity and requires changes to the underlying architecture.

2. Variability in Whisper AI's Accuracy:

The voice recognition performance of Whisper AI depends heavily on the quality of the audio input. The accuracy of transcription can be reduced when working with large environments or poor images. To mitigate this, the inclusion of noise reduction techniques or adaptive filtering can help improve transcription clarity and accuracy, and ensure more reliable language-based interactions.

3. Dependence on Mistral API Performance:

The responsiveness of the Mistral API-based text generation system depends on continuous development and optimization. If AI models develop quickly, the efficiency and maintenance of new progress requires continuous updates and improvements. The inclusion of real-time adaptability guarantees and fallback mechanisms to API is important to maintain system negligence.

ACKNOWLEDGMENTS

We would like to thank Dr. B. Bhanu Prakash sir Head of the Department(HOD), Dr. T. Syam Sundara Rao sir, and the Trainers Especially, a special thanks to Teja Kavali sir for their continuous support, encouragement, and guidance throughout the project. Their insight and leadership were crucial in defining our work. Their valuable feedback and suggestions helped us refine our approach and increase the overall effectiveness of our implementation. Their dedication and hard work have enabled us to understand AI, natural language processing, and the healing process. We thank you for your steadfast support and trust in our work.

VI. REFERENCES

[1] Lewis, P. P. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 9459-9474.

[2] FINE-TUNE THE ENTIRE RAG ARCHITECTURE (INCLUDING DPR RETRIEVER) FOR QUESTION-ANSWERING,**arXiv:2106.11517v1 [cs.IR] 22 Jun 2021**.

[3] Kelvin Guu, Kenton Lee, Zora Tung, PanupongPasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. ArXiv, abs/2002.08909, 2020.

[4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and HaofenWang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[5] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya,Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models.*arXiv preprint arXiv:2203.15556* (2022).

[6] OpenAI.2024.OpenAIAPIModels.https://platform.openai.com/docs/models/gpt-3-5-turbo. [Online; accessed 04-February-2024].

[7] Haoyan Luo and Lucia Specia. 2024. From Understanding to Utilization: A Survey on Explainability for Large Language Models. *arXiv preprint arXiv:2401.12874* (2024).

[8] OpenAI. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. *OpenAI Research*.

[9] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*

[10] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

[11] Mistral AI. (2023). "Mistral API Documentation." Retrieved from: https://mistral.ai

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.

[13] Reimers, N., &Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *EMNLP Conference Paper*.

[14] Radford, A., Kim, J. W., Hallacy, C., et al. (2022). "Robust Speech Recognition via Large-Scale Weak Supervision." *OpenAI Research Paper*.

[15] Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[16] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., & Yih, W.-t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. arXiv preprint arXiv:2004.04906.

[17] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.

[18] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[19] Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu.com.

[20] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed. draft). Prentice Hall.

[21] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[22] International Journal on "Wielding Neural Networks to Interpret Facial Emotions in Photographs with Fragmentary Occlusion", on American Scientific Publishing Group (ASPG) Fusion: Practice and Applications(FPA) ,Vol. 17, No. 01, August, 2024, pp. 146-158.

[23] International Journal on "Prediction of novel malware using hybrid convolution neural network and long short-term memory approach", on International Journal of Electrical and Computer Engineering (IJECE),Vol. 14, No. 04, August, 2024, pp. 4508-4517.

[24] International Journal on "Cross-Platform Malware Classification: Fusion of CNN and GRU Models",on International Journal of Safety and Security Engineering (IIETA),Vol. 14, No. 02, April, 2024, pp. 477-486

[25] International Journal on "Enhanced Malware Family Classification via Image-Based Analysis Utilizing a Balance-Augmented VGG16 Model,onInternational information and Engineering Technology Association (IIETA),Vol. 40, No. 5, October, 2023, pp. 2169-2178

[26] International Journal on "Android Malware Classification Using LSTM Model, International information and Engineering Technology Association (IIETA) Vol. 36, No. 5, (October, 2022), pp. 761 – 767. Android Malware Classification Using LSTM Model | IIETA.

[27] International Journal on "Classification of Image spam Using Convolution Neural Network", Traitement du Signal, Vol. 39, No. 1, (February 2022), pp. 363-369 .

[28] International Journal on "Medical Image Classification Using Deep Learning Based Hybrid Model with CNN and Encoder", International information and Engineering Technology Association (IIETA), Revue d'IntelligenceArtificiellVol. 34, No. 5, (October, 2020), pp. 645 – 652.

[29] International Journal on "Prediction of Hospital Re-admission Using Firefly Based Multi-layer Perception, International information and Engineering Technology Association (IIETA) Vol. 24, No. 4, (sept, 2020), pp. 527 – 533.

[30] International Journal on "Energy efficient intrusion detection using deep reinforcement learning approach",Journal of Green Engineering (JGE),Volume-11, Issue-1,January 2021.625-641.

[31] International Journal on "Classification of High Dimensional Class Imbalance Data Streams Using Improved Genetic Algorithm Sampling", International Journal of Advanced Science and Technology, Vol. 29, No. 5, (2020), pp. 5717 – 5726.

[32] Dr. M. Ayyappa Chakravarthi etal. published Springer paper "Machine Learning-Enhanced Self-Management for Energy-Effective and Secure Statistics Assortment in Unattended WSNs" in Springer Nature (Q1), Vol 6, Feb 4th 2025

[33] Dr. M. Ayyappa Chakravarthi etal. published Springer paper "GeoAgriGuard AI-Driven Pest and Disease Management with Remote Sensing for Global Food Security" in Springer Nature (Q1), Jan 20th 2025.

[34] Dr. M. Ayyappa Chakravarthi etal. presented and published IEEE paper "Machine Learning Algorithms for Automated Synthesis of Biocompatible Nanomaterials" , ISBN 979-8-3315-3995-5, Jan 2025.

[35] Dr. M. Ayyappa Chakravarthi etal. presented and published IEEE paper "Evolutionary Algorithms for Deep Learning in Secure Network Environments" ISBN:979-8-3315-3995-5, Jan 2025.

[36] Dr. Ayyappa Chakravarthi M. etal, published Scopus paper "Time Patient Monitoring Through Edge Computing: An IoT-Based Healthcare Architecture" in Frontiers in Health Informatics (FHI), Volume 13, Issue 3, ISSN-Online 2676-7104, 29th Nov 2024.

[37] Dr. Ayyappa Chakravarthi M. etal, published Scopus paper "Amalgamate Approaches Can Aid in the Early Detection of Coronary heart Disease" in Journal of Theoretical and Applied Information Technology (JATIT) , Volume 102, Issue 19, ISSN 1992-8645, 2nd Oct 2024.

[38] Dr. Ayyappa Chakravarthi M, etal, published Scopus paper "The BioShield Algorithm: Pioneering Real-Time Adaptive Security in IoT Networks through Nature-Inspired Machine Learning" in SSRG (Seventh Sense Research Group) -International Journal of Electrical and Electronics Engineering (IJEEE), Volume 11, Issue 9, ISSN 2348-8379, 28th Sept 2024.

[39] Ayyappa Chakravarthi M, Dr M. Thillaikarasi, Dr Bhanu Prakash Battula, published SCI paper "Classification of Image Spam Using Convolution Neural Network" in International Information and Engineering Technology Association (IIETA) - "Traitement du Signal" Volume 39, No. 1

[40] Ayyappa Chakravarthi M, Dr. M. Thillaikarasi, Dr. Bhanu Praksh Battula, published Scopus paper "Classification of Social Media Text Spam Using VAE-CNN and LSTM Model" in International Information and Engineering Technology Association (IIETA) - Ingénierie des Systèmesd'Information (Free Scopus) Volume 25, No. 6.

[41] Ayyappa Chakravarthi M, Dr. M. Thillaikarasi, Dr. Bhanu Praksh Battula, published Scopus paper "Social Media Text Data Classification using Enhanced TF_IDF based Feature Classification using Naive Bayesian Classifier" in International Journal of Advanced Science and Technology (IJAST) 2020

[42] Ayyappa Chakravarthi M. presented and published IEEE paper on "The Etymology of Bigdata on Government Processes" with DOI 10.1109/ICICES.2017.8070712 and is Scopus Indexed online in IEEE digital Xplore with Electronic ISBN: 978-1-5090-6135-8, Print on Demand(PoD) ISBN:978-1-5090-6136-5, Feb'2017.

[43] Subba Reddy Thumu& Geethanjali Nellore, Optimized Ensemble Support Vector Regression Models for Predicting Stock Prices with Multiple Kernels. Acta Informatica Pragensia, 13(1), x–x. 2024.

[44] Subba Reddy Thumu, Prof. N. Geethanjali. (2024). "Improving Cryptocurrency Price Prediction Accuracy with Multi-Kernel Support Vector Regression Approach". International Research Journal of Multidisciplinary Technovation 6 (4):20-31.

[45] Dr.Syamsundararaothalakola et.al. published Scopus paper "An Innovative Secure and Privacy-Preserving Federated Learning Based Hybrid Deep Learning Model for Intrusion Detection in Internet-Enabled Wireless Sensor Networks " in IEEE Transactions on Consumer Electronics 2024.

[46] Dr. Syamsundararaothalakola et.al. published Scopus paper "Securing Digital Records: A Synerigistic Approach with IoT and Blockchain for Enhanced Trust and Transparency " in International Journal of Intelligent Systems and Applications in Engineering 2024.

[47] Dr. Syamsundararaothalakola et.al. published Scopus paper "A Model for Safety Risk Evaluation of Connected Car Network " inReview of Computer Engineering Research2022.

[48] Dr. Syamsundararaothalakola et.al. published Scopus paper "An Efficient Signal Processing Algorithm for Detecting Abnormalities in EEG Signal Using CNN" in Contrast Media and Molecular Imaging 2022.