



International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)



ijerst.editor@gmail.com
editor@ijerst.com

Research Paper

MULTI - RAG ENTERPRISE

¹ A Satyanarayana, ² V Tejaswini, ³ S Bhavana, ⁴ G Vishnu, ⁵ M Sreeshanth

¹Assistant Professor, ^{2,3,4,5}Students

Department of CSE(AI ML)

Siddhartha Institute of Technology & Sciences, Narapally

drsatanarayanaakella_cse@siddhartha.co.in, 23TQ1A6660@siddhartha.co.in,

23TQ1A6635@siddhartha.co.in, 23TQ1A6630@siddhartha.co.in, 23TQ1A6658@siddhartha.co.in

ABSTRACT

Multi-RAG Enterprise is an intelligent enterprise knowledge-management and question-answering system that uses multiple Retrieval-Augmented Generation (RAG) pipelines to provide accurate and context-aware responses from organizational data. Modern enterprises store information across documents, databases, cloud storage, websites, emails, knowledge bases, and internal applications. Finding the right information from these distributed sources can be difficult and time-consuming.

The proposed system connects multiple enterprise data sources and creates specialized retrieval pipelines for different types of information. Documents can be processed, divided into meaningful chunks, converted into embeddings, and stored in vector databases for semantic retrieval. Structured information from relational databases and other systems can be accessed through appropriate query mechanisms rather than treating every source as unstructured text.

Multi-RAG Enterprise uses a routing and orchestration layer to determine which knowledge source or RAG pipeline is most relevant to a user's query. The system can retrieve information from multiple sources, combine relevant context, and provide a grounded response using a Generative AI model. Source references can be displayed with responses to improve transparency and help users verify the information.

The platform can support different enterprise departments such as Human Resources, Finance, IT, Sales, Legal, Operations, and Customer Support. Access-control policies can ensure that users retrieve only information they are authorized to access. The system can also maintain conversation context, query history, feedback, and retrieval performance for continuous improvement.

Overall, Multi-RAG Enterprise provides a centralized and intelligent approach to enterprise knowledge retrieval. It reduces information-search time, improves access to organizational knowledge, and supports data-driven decision-making. Future enhancements can include agentic retrieval, multimodal RAG, real-time data synchronization, advanced query planning, automatic source evaluation, enterprise workflow automation, and improved response-quality monitoring.

I. INTRODUCTION

Modern enterprises generate and maintain enormous amounts of information through business operations. Documents, policies, reports, emails, databases, project records,

technical manuals, and knowledge bases are continuously created and updated. Although this information is valuable, employees often face difficulties finding the exact information they need.

Traditional enterprise search systems generally depend on keyword matching, predefined filters, or manually organized categories. Such approaches may fail when users express questions in natural language or when the required information is distributed across multiple sources. Employees may need to search several applications separately before finding a complete answer.

Generative AI and Large Language Models have introduced new possibilities for enterprise knowledge systems. However, a language model operating without access to enterprise-specific information may produce outdated, incomplete, or unsupported responses. Retrieval-Augmented Generation addresses this problem by retrieving relevant information from trusted data sources and using that context to generate responses.

A single RAG pipeline may not be sufficient for a large enterprise because different departments and data sources have different structures, access policies, and retrieval requirements. **Multi-RAG Enterprise** therefore uses multiple specialized retrieval pipelines connected through an orchestration layer. The system can route a query to the appropriate source, retrieve relevant information, combine evidence, and generate a grounded response.

The objective of the proposed system is to provide a unified enterprise knowledge assistant that can understand natural-language queries while respecting organizational access controls. It can improve knowledge discovery, reduce repetitive information searches, and help employees access relevant organizational information more efficiently. The system should also provide source attribution and appropriate safeguards for sensitive enterprise information.

II. LITERATURE SURVEY

Traditional enterprise search systems use keyword-based information retrieval techniques to locate documents and records. These systems are useful when users know the exact terms contained in the source material. However, keyword-based approaches may struggle with natural-language questions, synonyms, context, and information distributed across multiple repositories.

Semantic search introduced embedding-based techniques that represent documents and queries as vectors. Vector databases can then retrieve information based on semantic similarity rather than exact keyword matching. This approach improves retrieval for natural-language questions but may still face challenges when information exists in different formats or when multiple enterprise sources must be combined.

Retrieval-Augmented Generation has emerged as an approach for combining information retrieval with Generative AI. Instead of relying only on information stored in a language model, RAG retrieves relevant external context and provides it to the model during response generation. This can improve grounding and allow systems to work with organization-specific and frequently updated information.

Enterprise RAG systems have increasingly incorporated document loaders, chunking strategies, embedding models, vector databases, reranking, metadata filtering, and access-control mechanisms. These components help improve retrieval quality and ensure that relevant information is provided to the generation model. However, different enterprise repositories may require different retrieval strategies.

Multi-RAG and multi-source retrieval architectures extend traditional RAG by using multiple knowledge stores or specialized retrieval pipelines. A query-routing or orchestration layer can determine which retrievers should be used for a particular question. Responses from multiple sources can then be combined and synthesized by a Generative AI model.

Recent intelligent knowledge systems are also exploring agentic RAG, hybrid search, knowledge graphs, multimodal retrieval, and automated evaluation. These approaches can improve complex enterprise question answering but introduce additional challenges related to latency, cost, data security, hallucination, source conflicts, and evaluation. The proposed **Multi-RAG Enterprise** addresses these challenges through specialized retrieval pipelines, intelligent routing, source-aware generation, access control, and response validation.

III. SYSTEM ANALYSIS

The **Multi-RAG Enterprise** system is designed to provide intelligent access to information distributed across multiple enterprise data sources. The system accepts natural-language questions through a unified enterprise chat interface. A query-understanding module analyzes the question to identify its intent, entities, department, and required information type. A RAG orchestration layer determines which specialized retrieval pipelines should process the query. Document-based RAG pipelines retrieve information from enterprise documents, policies, reports, and manuals. Structured-data pipelines retrieve information from authorized databases using appropriate query mechanisms. Additional connectors can access approved knowledge bases, cloud repositories, internal applications, and other enterprise sources. Retrieved information is filtered using metadata, relevance scores, and user-access permissions. A reranking module can prioritize the most relevant retrieved content before generation. The Generative AI engine combines the selected context and produces a grounded natural-language response. A validation layer checks source availability, response relevance, and potential unsupported claims. Source references can be displayed so users can verify the information used to generate the response. Finally, user feedback and retrieval-performance metrics can be collected to improve the system over time.

Existing System

Existing enterprise knowledge systems commonly use separate search interfaces for documents, databases, emails, cloud storage, knowledge bases, and departmental applications. Employees may need to access multiple systems to find information related to a single business question. Traditional keyword search may return large numbers of documents without understanding the actual intent of the user. Some enterprise systems provide limited semantic search but may focus on a single repository. Standard chatbot systems may provide conversational interfaces but can lack access to

current internal information. General-purpose Generative AI systems may not have access to confidential enterprise data and may generate unsupported responses when information is unavailable. Separate departmental systems can also create information silos across organizations. Access permissions may be implemented differently across different applications. Combining information from multiple sources may therefore require manual effort. Existing systems may provide limited source attribution and cross-source reasoning. They may also lack centralized monitoring of retrieval quality and response accuracy. These limitations create a need for an integrated multi-source enterprise RAG platform.

Disadvantages of Existing System

- Enterprise information is distributed across multiple applications.
- Employees may need to search several systems separately.
- Keyword search may fail to understand user intent.
- Limited semantic understanding in traditional search.
- Single-source RAG systems have limited enterprise coverage.
- General AI systems may lack organization-specific knowledge.
- Risk of unsupported or hallucinated responses.
- Difficult cross-source information retrieval.
- Different systems may have different access-control policies.
- Limited source attribution and verification.
- Manual information gathering consumes time.
- Limited centralized monitoring of retrieval quality.

Proposed System

The proposed **Multi-RAG Enterprise** provides a unified AI-powered knowledge platform that connects multiple enterprise information sources through specialized retrieval pipelines. The system accepts natural-language questions through a centralized conversational interface and analyzes each query to determine its intent and required knowledge domains. An intelligent orchestration layer routes the query to one or more appropriate RAG pipelines based on the data required. Unstructured document RAG retrieves relevant content from policies, manuals, reports, and enterprise documents. Structured-data retrieval connects to authorized databases and business systems using appropriate query methods. Additional enterprise connectors can retrieve information from approved cloud repositories, knowledge bases, and internal applications. Retrieved content is filtered according to relevance, metadata, and user authorization. A reranking component selects the most useful evidence before passing it to the Generative AI engine. The generation module creates a concise response grounded in the retrieved enterprise information. Source references are provided to improve transparency and verification. A response-validation layer identifies potential unsupported responses or insufficient evidence. The system can collect user feedback and performance metrics to improve routing and retrieval quality.

Advantages of Proposed System

- Provides a unified enterprise knowledge interface.
- Connects multiple organizational data sources.
- Supports multiple specialized RAG pipelines.

- Provides intelligent query routing.
- Supports both structured and unstructured data.
- Improves semantic information retrieval.
- Reduces time spent searching for information.
- Provides grounded Generative AI responses.
- Supports source attribution and verification.
- Applies user-level access controls.
- Supports cross-department knowledge discovery.
- Enables retrieval and response-quality monitoring.

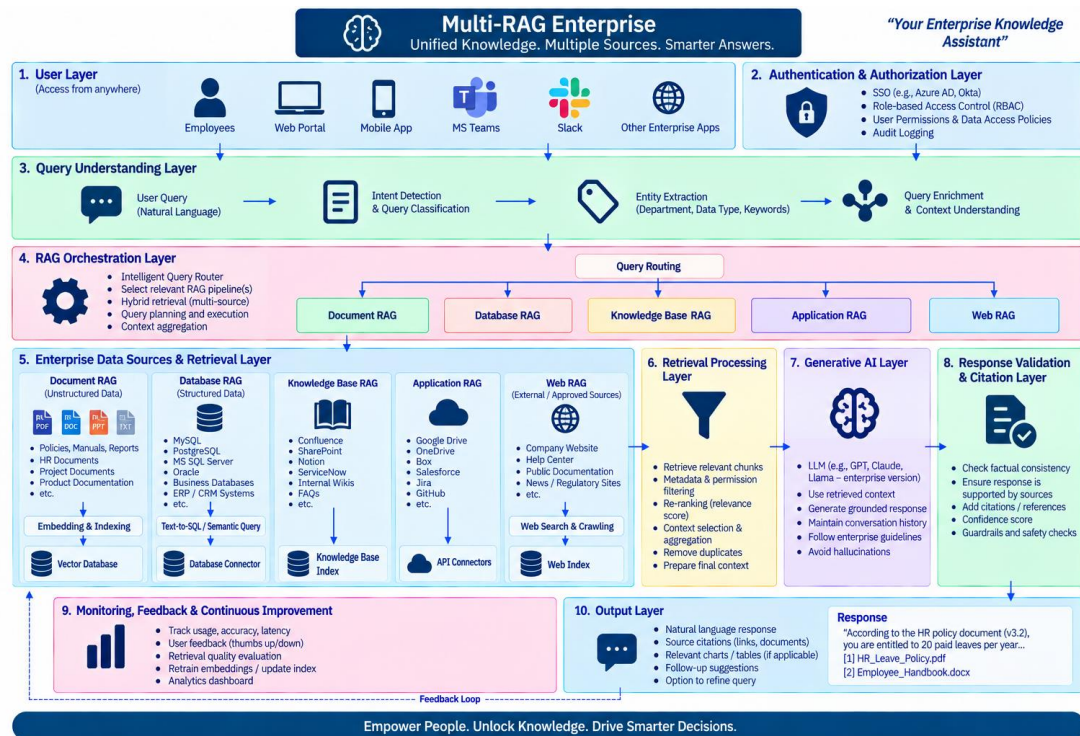
IV. METHODOLOGY

The methodology begins when an authorized enterprise user submits a natural-language question through the conversational interface. The query-processing module analyzes the question and identifies its intent, entities, department, and required information. The orchestration layer determines which RAG pipelines and enterprise sources are relevant to the query. Unstructured document pipelines preprocess documents through parsing, chunking, metadata extraction, and embedding generation. Embeddings are stored in appropriate vector databases to support semantic retrieval. Structured-data pipelines connect to approved databases and generate suitable queries for retrieving relevant records. The retrieval layer collects candidate information from one or more sources while applying user permissions and metadata filters. A reranking module evaluates retrieved content and prioritizes the most relevant evidence. The selected context is passed to the Generative AI model along with instructions to generate a grounded response. A response-validation module checks whether the answer is supported by the retrieved evidence and whether source references are available. The final response is displayed through the enterprise chat interface together with relevant citations or source information. User feedback, retrieval metrics, latency, and response-quality measurements are collected for system evaluation. Finally, monitoring and continuous improvement mechanisms are used to optimize retrieval strategies, routing, and response quality.

SYSTEM ARCHITECTURE

The Multi-RAG Enterprise architecture consists of multiple interconnected layers that enable secure retrieval and generation from distributed enterprise knowledge sources. The User Interface Layer provides a unified conversational interface for authorized employees. The Authentication and Authorization Layer verifies user identity and determines which enterprise information the user can access. The Query Understanding Layer analyzes natural-language questions and identifies intent, entities, and required knowledge domains. The RAG Orchestration Layer routes each query to one or more appropriate retrieval pipelines. The Document RAG Layer processes and retrieves information from enterprise documents, policies, manuals, reports, and knowledge bases. The Structured Data RAG Layer connects to authorized databases and business systems for structured information retrieval. Additional Enterprise Connectors provide access to approved cloud repositories, internal applications, and other organizational sources. The Retrieval and Reranking Layer combines candidate results, applies metadata and permission filters, and selects the most relevant evidence. The Generative AI Layer uses the retrieved context to generate a grounded response. The Validation

and Citation Layer checks response support, source availability, and potential unsupported claims.



V. RESULT AND OUTPUT

Multi-RAG Enterprise
One Query. Multiple Sources. Trusted Answers.

Multi-RAG Enterprise
Ask. Discover. Decide.

Search knowledge, documents, people...

Arjun Sharma
Product Team

What is our company's work from home policy?

10:24 AM

According to the HR Policy Document (v3.2), our company follows a hybrid work model.

- Employees can work from home up to 3 days per week with manager approval.
- Core working hours (10 AM - 4 PM) are mandatory.
- Work from home requests should be raised through the HR portal.
- Certain roles (e.g., security, operations) are not eligible for remote work.
- For more details, please refer to the official policy document.

10:24 AM

Sources

- HR_Work_From_Home_Policy.pdf
HR Portal • Updated: Jan 2024
[View Source](#)
- Employee_Handbook.docx
Intranet • Updated: Dec 2023
[View Source](#)
- Remote_Work_FAQ
Confluence • Updated: Feb 2024
[View Source](#)

Ask a question about your company's data...

Summarize this document | Compare with previous policy | Show applicable to my role | Give key points

Retrieved from Multiple Sources

- HR Documents
1 document used
- Knowledge Base
1 article used
- Company Intranet
1 page used

Answer Confidence

92% **High Confidence**
Answer is based on verified company sources.

Related Questions

- How to apply for work from home?
- Is work from home allowed for interns?
- What are the core working hours?
- Who approves remote work requests?

More Example Results

- HR** Policies, Leaves, Benefits
What is the maternity leave policy?
Maternity leave is 26 weeks for eligible employees as per HR Policy (v3.1)...
[Maternity_Leave_Policy.pdf](#)
[View Source](#)
- Finance** Budgets, Reports, Expenses
What was the Q2 revenue?
The total revenue for Q2 2024 was ₹ 12.5 Crores, which is a 12% increase compared to Q1 2024.
[Q2_Financial_Report.xlsx](#)
[View Source](#)
- IT** Support, Guides, Troubleshooting
How to reset my VPN password?
You can reset your VPN password using the self-service portal. Follow the steps in the IT Guide.
[VPN_Reset_Guide.pdf](#)
[View Source](#)
- Sales** Reports, Leads, Strategies
Show top 5 clients by revenue.

Rank	Client Name	Revenue (₹ Cr)
1	ABC Corp	5.2
2	TechNova	3.8
3	GlobalMart	3.1
4	HealthPlus	2.9
5	NextGen	2.4
- Legal** Contracts, Compliance
What is the NDA policy?
All employees and external partners must sign the NDA before accessing confidential information...
[NDA_Policy.pdf](#)
[View Source](#)

50K+ Documents Indexed | 10+ Data Sources | 5K+ Active Users | 70% Faster Information Access

Knowledge across your enterprise. Answers in seconds

VI. CONCLUSION

The Multi-RAG Enterprise system provides an intelligent and centralized solution for accessing information distributed across different enterprise data sources. By combining multiple Retrieval-Augmented Generation pipelines, the system allows employees to interact with organizational knowledge through a single natural-language interface instead of searching through multiple applications separately.

The system combines Natural Language Processing, intelligent query routing, semantic retrieval, vector databases, structured database retrieval, Generative AI, and response validation to produce context-aware and grounded answers. The orchestration layer identifies the appropriate knowledge sources, retrieves relevant information, and provides the selected context to the Generative AI model for response generation.

One of the major advantages of the system is its ability to work with different types of enterprise information, including documents, databases, knowledge bases, cloud repositories, and internal applications. Source references and response-validation mechanisms improve transparency and allow users to verify the information used to generate an answer.

Overall, Multi-RAG Enterprise can reduce information-search time, improve enterprise knowledge accessibility, and support more efficient decision-making. It provides a scalable foundation for departments such as HR, Finance, IT, Sales, Legal, and Operations while maintaining appropriate access-control mechanisms.

Future enhancements can include agentic RAG, multimodal document retrieval, real-time data synchronization, knowledge-graph integration, advanced query planning, automated source-quality evaluation, multilingual support, and enterprise workflow automation. Strong authentication, authorization, data isolation, privacy controls, monitoring, and human oversight should remain core requirements for secure and reliable enterprise deployment.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, 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.
- [1] 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.
- [2] 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.
- [3] 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.
- [4] M. M. Rao, M. B. Mohiuddin, P. Srinu, A. Satyanarayana, J. Madhavan and R. D. Kumar, "Carbon-Neutral-Agriculture: Renewable-Energy Powered IoT Systems and Resource Optimization," 2026 3rd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2026, pp. 1-6, doi: 10.1109/RMKMATE69073.2026.11518903.
- [5] R. Uma, M. Maggidi, A. R. Ekkati, S. Bandlamudi, C. Madugundi and S. K. Taduri, "Zero-Trust Architecture for Hybrid Cloud Security in Critical Infrastructures," 2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON), Lonavala, India, 2026, pp. 1-6, doi:10.1109/I3CTCON68242.2026.11507458.
- [6] VNS Manaswini, Chithanoor Arjun, Yamini Chouhan, & V. Sumalatha. (2026). MB-EFF: A Multi-Branch Environmental Feature Fusion Deep Learning Model for Crop Classification and Recommendation. *International Journal of Computer*

Information Systems and Industrial Management Applications, 18(22s), 460–468.

<https://doi.org/10.70917/ijcisim-2026-5453>

[7] Yamini Chouhan, SIDDOJU DIVYA, VNS Manaswini, & J. Priyanka. (2026). Vision-Threatening Diabetic Retinopathy Detection Using EfficientNetB4: A Two-Stage Transfer Learning Approach. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(22s), 469–476.

<https://doi.org/10.70917/ijcisim-2026-5454>

[8] Real Time Vehicle Counting and Classification system, **Dr. A.Satyanarayana,et.al** International Research Journal of Modernization in Engineering Technology and Science, Vol-06, Issue-05, 2024 [ISSN: **2582-5208**] doi: 10.56726/IRJMETS57804

[9] Facial Depth Maps: Detecting Sleep Apnea with Deep Learning, **Dr. A.Satyanarayana,et.al** , International Journal of Scientific Development and Research, Vol 9, Issue-5, 2024 [ISSN: **2455-2631**]

[10] Moodchef: Mood Based Smart Recipe Recommender with cooking Assistance, **Dr. A.Satyanarayana,et.al** , International Journal of Progressive Research in Engineering Management and Sciences, Vol-05,Issue-7, July 2025 [ISSN: **2583-5162**] doi: **10.58257/IJPREMS42696**

[11] News Article Text Summarization, **Dr. A.Satyanarayana,et.al**, Journal of Emerging Technologies and Innovative Research, Vol-11, Issue -5, 2024, [ISSN: **2349-5162**]

[12] Voice Assistant and gesture controlled virtual mouse using Deep learning Techniques, **Dr. A.Satyanarayana,et.al** , International Research Journal of Modernization in Engineering Technology and Science Vol-6, Issue-5, 2024 [ISSN: **2582-5208**] **10.56726/IRJMETS57214**