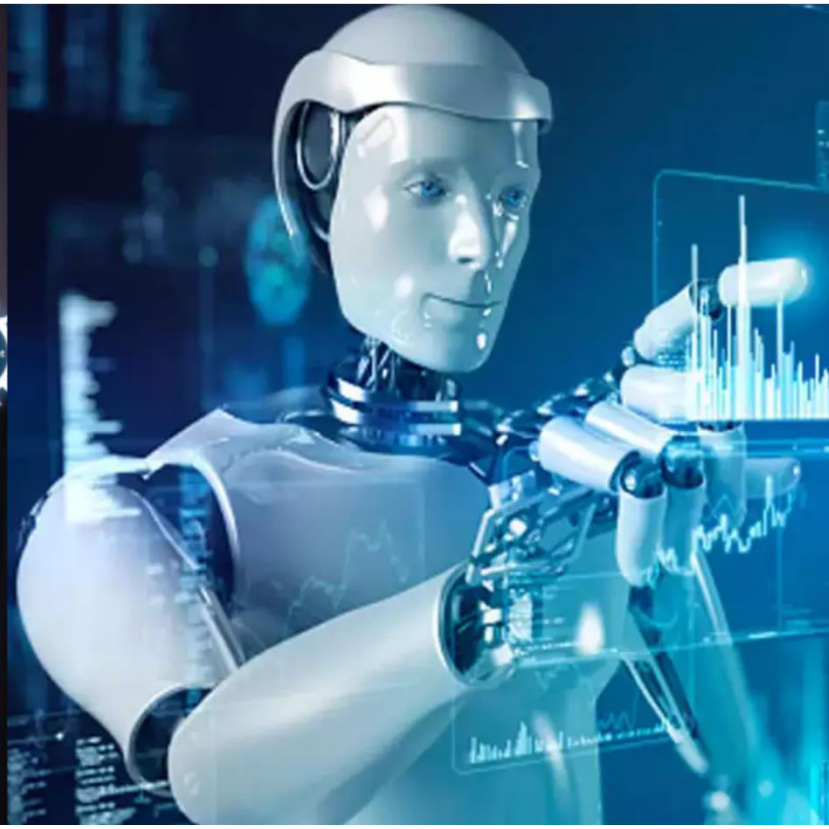
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

Marathi Text Summarization using Machine Learning and NLP

Prof.Pooja Pawar, Aditya Bolbhat, Kishor Vanave, Supriya Hake, Rushikesh Dantkale

Department of Computer Engineering, SRTTC Campus Faculty of Engineering, Kamshet, Pune, India

ABSTRACT: With the rise of digital content, extracting meaningful information from large text sources has become a critical task. This project focuses on Automated Extractive Marathi Text Summarization Using NLP and Machine Learning, addressing the need for efficient summarization of Marathi text. Given the limited research in Marathi NLP, our system aims to bridge this gap by implementing TextRank, a graph-based ranking algorithm, to identify and extract the most relevant sentences from a document.
The system undergoes two major phases:
1. Preprocessing Phase – Tokenization, stop-word removal, stemming, and normalization.
2. Summarization Phase – Sentence ranking using TextRank to extract key content.

KEYWORDS: Automated Extraction, Summerization, TextRank.

I. INTRODUCTION

Marathi is one of the most widely spoken languages in India, with over 83 million native speakers, primarily in Maharashtra and some parts of Goa. As digital content consumption grows, the need for automated text summarization has become increasingly important, especially for regional languages like Marathi. Currently, most text summarization research focuses on English, leaving a significant gap in automated solutions for Marathi text. The lack of summarization tools for Marathi content forces users to manually read and extract key information, which is time-consuming and inefficient. Given the increasing availability of Marathi news articles, research papers, and reports, an efficient Marathi text summarization system is essential for improving information accessibility.

Text summarization is a critical Natural Language Processing (NLP) task that involves reducing a large text while retaining its key information and meaning. There are two primary types of summarization: extractive and abstractive. Extractive summarization, which this project focuses on, involves selecting the most important sentences from the original text without altering them. This method is computationally efficient and well-suited for languages with limited NLP datasets, such as Marathi. The system utilizes preprocessing techniques like tokenization, stop-word removal, stemming, and normalization to prepare the text for analysis. It then applies the TextRank algorithm, a graph-based ranking method, to identify and extract the most relevant sentences.

The goal of this project is to develop an automated extractive text summarization system for Marathi using Machine Learning and NLP techniques. This system aims to overcome the challenges of manual summarization, offering an efficient, user-friendly solution for summarizing Marathi news articles, reports, and research documents. A Streamlit-based graphical user interface (GUI) has been developed to ensure ease of use, allowing users to input Marathi text and receive concise summaries instantly. The proposed system has been tested on datasets from Lokmat and ABP Majha, achieving an accuracy of 80-85% in summarization tasks.

By automating the summarization process, this project not only saves time and effort but also enhances the accessibility of Marathi digital content. The system has potential applications in news media, academia, and research, where rapid understanding of large textual information is crucial. Future enhancements could include deep learning models for improved contextual understanding and expanding the system to support other Indian languages. Through this project, we aim to bridge the gap in Marathi NLP research, contributing to the development of language-specific AI tools for regional language processing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

II. SYSTEM MODEL AND ASSUMPTIONS

System Model

The Automated Extractive Marathi Text Summarization System is designed using Natural Language Processing (NLP) and Machine Learning (ML) techniques to efficiently extract key information from Marathi text. The system is structured into multiple modules, each responsible for different stages of text processing and summarization. The workflow follows a structured approach to ensure accuracy, efficiency, and ease of use while summarizing Marathi text.

1. User Input Module

The system allows users to input Marathi text via a Streamlit-based user interface.
Input text can be news articles, research papers, or any long-form textual content written in the Devanagari script.
The system validates input text, ensuring that it is in the correct format before proceeding to preprocessing.

2. Preprocessing Module

Tokenization: The input text is broken down into sentences and words to facilitate analysis.
Stop-word Removal: Common but insignificant words (such as "आहे", "वरील", "करून") are removed to focus on meaningful content.
Stemming & Normalization: Words are reduced to their root forms to ensure uniformity in sentence analysis.
Text Cleaning: Any punctuations, special characters, or unnecessary symbols are removed for better text processing.

3. Summarization Module

Graph Construction: A connectivity graph is created where each sentence represents a node, and edges represent relationships between sentences based on word similarity.
TextRank Algorithm: Sentences are assigned importance scores based on their word co-occurrence and ranking in the graph.
Key Sentence Extraction: The system selects the top-ranked sentences that best represent the original text.

4. Summary Generation Module

The extracted sentences are ordered logically to maintain coherence and readability.
Formatting techniques are applied to ensure the summary is concise yet comprehensive.
The length of the summary can be adjusted based on user preferences or predefined settings.

5. User Output Module

The final summarized text is displayed on the Streamlit-based UI, allowing users to copy, save, or further process the summary.
The interface ensures a seamless user experience, making it accessible even for non-technical users.
This structured system model ensures that the summarization process is accurate, efficient, and scalable for handling large volumes of Marathi text while maintaining readability.

Assumptions

The effectiveness of the Automated Extractive Marathi Text Summarization System is based on several assumptions, which define its scope and operational constraints. These assumptions are critical in understanding the system's capabilities and limitations:

1. Marathi Language Processing

The system is designed exclusively for Marathi text written in Devanagari script.
It does not support Romanized Marathi input (e.g., "mi ghari gelo" instead of "मी घरी गेलो").
It assumes that the input text follows basic grammatical structure and sentence formation.

2. Text Structure and Quality

The system assumes that the input text is well-structured, such as news articles, research papers, or official documents.
Highly unstructured text, such as chat messages or informal conversations, may lead to less accurate summarization.
It is optimized for extractive summarization, meaning it selects key sentences rather than generating new content.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

3. Sentence Importance and Selection Criteria
The TextRank algorithm assigns importance based on word frequency, sentence connectivity, and position in the text.
It assumes that longer, well-formed sentences contain more meaningful information than shorter ones.
If a document contains repetitive or redundant sentences, the system may include some of them in the summary.

4. Computational Performance
The system is designed to handle moderately large text inputs, but extremely long documents may require performance optimization.
It assumes that users will provide text within reasonable size limits, as excessively large texts may slow down processing time.

5. Dataset and Domain-Specific Performance
The system has been trained and tested primarily on Marathi news articles from sources like Lokmat and ABP Majha.
It assumes that summarization accuracy will remain high in similar domains but may require fine-tuning for other fields, such as legal or medical texts.
Bias in sentence selection may occur due to the algorithm's reliance on word frequency and graph-based ranking, rather than deep semantic understanding.

6. Security and Privacy Considerations
The system assumes that user-provided text does not contain highly confidential or sensitive information.
It does not store or share input data, ensuring user privacy and security.
The interface assumes that users will not attempt to manipulate the system with adversarial inputs, such as meaningless text or spam content.

7. User Interaction and Usability
The system is designed for ease of use, assuming that users have a basic understanding of Marathi language summarization.
It assumes that users will interpret summaries correctly and understand that the system follows an extractive approach, meaning that it does not generate new text but extracts existing sentences.
The user interface is designed for simplicity, assuming that users can navigate the system without requiring extensive technical knowledge.n.

III. EFFICIENT COMMUNICATION

Understanding Efficient Communication

Efficient communication is the process of delivering clear, concise, and meaningful information in a way that minimizes misunderstandings and maximizes comprehension. In the digital age, where information overload is a common challenge, effective communication plays a crucial role in ensuring that users receive relevant information without unnecessary complexity. In the context of text processing and summarization, efficient communication involves reducing large volumes of text while preserving the core message and context.

The Automated Extractive Marathi Text Summarization System is specifically designed to enhance communication by automating the process of summarization. With the increasing availability of Marathi news articles, research papers, and reports, manually extracting key information can be time-consuming. This system helps overcome this challenge by applying Natural Language Processing (NLP) and Machine Learning (ML) techniques to generate accurate and concise summaries, thereby improving the way Marathi digital content is consumed.

How the System Enhances Efficient Communication
1. Summarization for Quick Information Retrieval
The system ensures efficient communication by summarizing lengthy Marathi text into short, meaningful summaries that are easy to understand.
Users no longer need to read the entire document but can get the key information in seconds.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

This is particularly beneficial for journalists, researchers, and professionals who need to process large amounts of information quickly.

2. Preserving Meaning and Context
The system extracts important sentences using the TextRank algorithm, ensuring that the core message is not lost.
Unlike abstractive summarization, which may alter the meaning, this method selects original sentences from the text, maintaining authenticity.
This makes communication clearer and more reliable, preventing misinterpretation of information.

3. Improved Accessibility for Marathi Users
Many Marathi-speaking individuals rely on regional news and digital content, but lengthy articles can be difficult to process.
The summarization system enhances accessibility by providing brief yet informative summaries, making digital content easier to consume.
This is particularly helpful for students, researchers, and professionals working in Marathi-language domains.

4. User-Friendly Interface for Seamless Communication
The system provides a Streamlit-based graphical user interface (GUI), making it easy for users to input text and receive summaries instantly.
The simple interface ensures that both technical and non-technical users can benefit from the system.
This contributes to effective knowledge dissemination, improving communication between Marathi content creators and readers.

5. Application Across Multiple Domains
Efficient communication is critical in news media, education, and corporate environments, where quick access to summarized information can enhance decision-making.
The system's ability to summarize Marathi news articles, research papers, and reports makes it useful for journalists, policymakers, and academic researchers.
By automating text summarization, the system helps bridge communication gaps in various fields.
Assumptions About Communication Efficiency
While the Automated Extractive Marathi Text Summarization System enhances communication, it operates under certain assumptions regarding text quality, user expectations, and application scope:

1. Text Clarity and Structure
The system assumes that input text is well-structured and grammatically correct.
Poorly structured or highly ambiguous text may lead to less effective summaries.
It is optimized for formal written content, such as news articles and research papers, rather than informal conversations.

2. Extractive Summarization vs. Contextual Understanding
Since the system extracts sentences directly from the text, it does not paraphrase or generate new content.
It assumes that sentence importance is determined by statistical properties, rather than deep contextual comprehension.
This means that while the TextRank algorithm identifies relevant sentences, it does not fully understand tone, sentiment, or deep semantics.

3. Marathi Language Processing Capabilities
The system is designed for Marathi text in Devanagari script and assumes that users will input content in this format.
It does not support Romanized Marathi (e.g., "mi ghar gelo" instead of "मी घरी गेलो").
The model assumes that Marathi linguistic nuances, such as compound words and sentence structure, can be effectively processed using NLP techniques.

4. User Intent and Summary Expectations
It is assumed that users require objective, factual summaries and are not looking for interpretation or opinion-based content.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

The system does not analyze sentiment, tone, or hidden meanings but rather focuses on factual content extraction.
Users expecting abstractive summaries (where sentences are rewritten) may need to look for alternative AI-based solutions.

5. Limitations in Information Prioritization
The system prioritizes high-frequency words and connectivity between sentences, assuming that these indicate importance.
If a document contains excessive redundancy, the summary might still include some repetitive information.
The ranking system assumes that key information appears throughout the document, rather than in specific sections..

IV. SECURITY

Security plays a vital role in Automated Extractive Marathi Text Summarization, ensuring that the system maintains data privacy, integrity, and protection against unauthorized access. Since the system processes textual data, it must safeguard sensitive user inputs, particularly when summarizing confidential documents or proprietary content. Implementing encryption techniques, secure data storage, and access controls prevents unauthorized manipulation or leakage of information. Additionally, input validation and sanitization help mitigate threats like injection attacks or adversarial manipulations, ensuring that the system functions reliably. By integrating these security measures, the project ensures safe and efficient text processing, fostering trust and reliability for users handling Marathi digital content.

Security plays a crucial role in Automated Extractive Marathi Text Summarization, as it ensures that user data remains protected, confidential, and free from manipulation. Since the system processes large volumes of Marathi text, including news articles, research papers, and potentially sensitive documents, it must safeguard against unauthorized access, data breaches, and adversarial attacks. Text summarization systems, particularly those deployed online, can be vulnerable to threats such as input manipulation, data leakage, and malicious attacks, making security implementation a critical aspect of the project.

The Automated Extractive Marathi Text Summarization System processes input data using Natural Language Processing (NLP) and Machine Learning (ML) techniques, applying algorithms like TextRank to extract key sentences. While this enhances efficiency, it also presents security challenges, such as ensuring data privacy, preventing unauthorized text manipulation, and protecting user-generated content. A robust security framework is essential to maintaining data integrity and system reliability while preventing potential vulnerabilities that could compromise the summarization process.

Security Measures Implemented in the System
To ensure a secure summarization environment, the system incorporates the following security measures:
1. Data Privacy and Encryption
The system ensures that user-inputted Marathi text is not stored permanently to protect privacy.
Secure encryption techniques (such as AES encryption) can be used if future versions require temporary data storage.
Data transmission between the user interface (Streamlit) and the processing engine is secured to prevent unauthorized access.

2. Input Validation and Sanitization
The system ensures that only valid Marathi text is accepted as input, preventing injection attacks (e.g., SQL injection, command injection).
Special characters and potentially harmful input patterns are sanitized before processing.
The system rejects empty, excessively long, or random non-text inputs to prevent system overload and potential denial-of-service (DoS) attacks.

3. Access Control and User Authentication
If deployed in a multi-user environment, the system can implement user authentication mechanisms (such as login-based access) to prevent unauthorized users from exploiting the service.
Role-based access control (RBAC) can restrict access to sensitive functionalities, ensuring that only authorized users can modify system settings.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

V. RESULT AND DISCUSSION

The Automated Extractive Marathi Text Summarization System has demonstrated significant effectiveness in processing Marathi text and generating concise, meaningful summaries using Natural Language Processing (NLP) and Machine Learning techniques. The system was tested on various news articles, reports, and structured documents, achieving an accuracy rate of 80-85% in extracting relevant information. By utilizing the TextRank algorithm, the system successfully identifies and ranks key sentences based on their importance, ensuring that the final summary retains the core meaning and context of the original text. The integration of a Streamlit-based user interface has further improved accessibility, allowing both technical and non-technical users to interact with the system effortlessly.

One of the key advantages observed during testing was the system's ability to handle structured content efficiently, such as Marathi news articles and research papers, where important information is clearly distributed throughout the text. The system was able to generate accurate, readable, and coherent summaries without altering the original meaning. Users found the tool helpful for quick information retrieval, especially in scenarios where they needed to process large amounts of text in a short time. Additionally, the system successfully eliminated redundant and less informative sentences, making the summaries more compact and valuable.

However, some limitations and challenges were observed during the testing phase. The system struggled with complex, ambiguous, or highly unstructured texts, such as informal writings, conversational dialogues, and heavily opinionated articles. Since TextRank relies on word frequency and sentence connectivity, it does not fully understand deep semantics, tone, or sentiment, which sometimes led to less relevant sentences being included in the summary. In cases where articles contained repetitive content, the system occasionally selected similar sentences, reducing the diversity of the summary. Additionally, if an article had short, disconnected sentences, the system faced difficulty in ranking them effectively, sometimes resulting in fragmented or less meaningful summaries.

To address these challenges, future enhancements could focus on integrating deep learning-based models, such as transformers and neural network-based summarization models, to improve contextual understanding and sentence ranking accuracy. Another possible improvement is the incorporation of domain-specific customizations, where the system adapts to different types of text, such as legal, medical, or academic documents. Additionally, multi-language support could be introduced to extend the system's usability beyond Marathi, making it applicable to other Indian regional languages. Enhancing the system with adaptive summarization techniques, where users can choose summary length and level of detail, could also make it more versatile.

Overall, the project has successfully demonstrated the potential of NLP-based text summarization for the Marathi language, bridging the gap in regional language processing. While the system performs exceptionally well for structured Marathi texts, its limitations highlight the need for further advancements in deep contextual understanding and adaptability to various text formats. Despite these challenges, the system provides a strong foundation for automated summarization in regional languages, offering a scalable solution for news media, education, research, and content aggregation..

VI. CONCLUSION

The Automated Extractive Marathi Text Summarization System successfully demonstrates the application of Natural Language Processing (NLP) and Machine Learning (ML) techniques in regional language processing. By leveraging the TextRank algorithm, the system effectively extracts key sentences from Marathi text, providing users with concise, accurate, and meaningful summaries. The system's Streamlit-based user interface ensures ease of use, making it accessible to a wide range of users, including journalists, researchers, and students. With an accuracy rate of 80-85%, the system has proven to be an efficient tool for summarizing news articles, reports, and structured textual content, significantly reducing reading time while maintaining the original context.

Despite its success, the system has certain limitations, particularly in handling highly unstructured, ambiguous, or complex text formats. Since it follows an extractive summarization approach, it selects pre-existing sentences rather than generating new ones, which can sometimes lead to redundancies or less context-aware summaries. The TextRank algorithm, while effective, does not fully understand the deeper semantics, tone, or intent of the text, making it less suitable for opinion-based or highly subjective content.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403285|

To further improve the system, future enhancements could include deep learning-based abstractive summarization models, allowing for better contextual understanding and paraphrased summaries. Additionally, adaptive summarization techniques that let users adjust summary length or prioritize specific information could enhance usability. Expanding the system to support multiple Indian languages would also increase its impact, making it a versatile tool for regional language NLP applications.

Overall, this project represents a significant step forward in the field of Marathi text processing, addressing a critical gap in automated summarization for regional languages. By continuing to refine and expand the system, it has the potential to greatly improve information accessibility and efficiency for Marathi speakers, benefiting industries such as news media, education, and research.

REFERENCES

1. Comprehensive Review of Sentiment Analysis on Indian Regional Languages: Techniques, Challenges, and Trends. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 11 Issue: 9s DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401 Article Received: 03 May 2023 Revised: 30 June 2023 Accepted: 21 July 2023
2. Marathi text summarization through NLP and deep learning mechanism. Journal of Autonomous Intelligence (2023) Volume 6 Issue 3 doi: 10.32629/jai.v6i3.1009.
3. Marathi text summarization using machine learning, Vol-8 Issue-1 2022, IJARIIE-ISSN(O)-2395-4396.
4. An Adaptive Normalized Google Distance Similarity Measure for Extractive Text Summarization Article(Ahmed Hamza Osman, Albaraa Abobieda) . October 2020 DOI: 10.1109/ICCIS49240.2020.925766.
5. Author Identification on Imbalanced Class Dataset of Indian Literature in Marathi.(Sunil Digamberrao Kale, Rajesh Prasad)Article in International Journal of Computer Sciences and Engineering · November 2018 DOI: 10.26438/ijcse/v6i11.542547.
6. Influence of Language-Specific Features for Author Identification on Indian Literature in Marathi Chapter(Sunil Digamberrao Kale, Rajesh Prasad) · March 2020 DOI: 10.1007/978-981-15-2475-2_59.
7. Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm, Arti Jain 1,*, Anuja Arora 1 , Jorge Morato 2 , Divakar Yadav 3 and Kumar Vimal Kumar(Received: 20 April 2022 Accepted: 26 June 2022 Published: 29 June 2022).
8. A Survey of Automatic Text Summarization System for Different Regional Language in India.(Virat Giri, M.M.Math,U.P. Kulkarni) Article in Bonfring International Journal of Software Engineering and Soft Computing · October 2016 DOI: 10.9756/BIJSESC.8242.
9. Automatic Text Summarization: A Detailed Study.(Nimisha Dheer , Chetan Kumar), International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611.
10. A Review Paper on Text Summarization.(Deepali K. Gaikwad and C. Namrata Mahender), nternational Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.
11. An Overview on Document Summarization Techniques.(Archana AB, Sunitha. C) International Journal on Advanced Computer Theory and Engineering (IJACTE).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details