# An Extensive Survey on Investigation Methodologies for Text Summarization

**Aahana Saklecha, Pragya Uplavdiya, M.P.S. Chawla**

***Abstract:*** *Natural language processing (NLP) is a fast-expanding field, and text summarization has recently gained alot of research interest. The necessity for automatic summarizing approaches to effectively digest massive amounts of textual data has grown in importance, due to the plethora (excessive amount of something) of information available in the digital age [18]. By automatically producing succinct and educational summaries of extensive materials, NLP-based text summarizing systems have the potential to revolutionize the way humans consume and process information. This review paper offers a thorough examination of the text summarizing research approaches. The process of creating a concise and useful summary of a text document is called text summarization. Evenfor cutting-edge natural language processing (NLP) systems, it is a difficult task. It was carried out using a thorough analysis of the most recent text summarizing research. The evaluation revealed a variety of research approaches that have been employed in the creation and assessment of text summarizing systems. This study's key discovery is that there are numerous different investigative approaches that can be used for text summarizing. These methods can be roughly divided into two groups:*

- *Extractive text summarization*
- *Abstractive text summarization*

*During the review we found that extractive summarization is a fairly simple method as it selects the key phrases from a text and extracts them to create a summary while abstractive summarization presents data in a clearer, more informative fashion by producing a summary. This review was important because it gives a thorough overview of the research approaches utilized for text summarizing, this article is significant. Researchers and programmers can utilize this data to create brand-new, improved text summarizing systems. [20]*

***Keyword:*** *Text Summarization, Natural Language Processing, Investigative Approaches*

## I. INTRODUCTION

The difficulty of choosing and categorizing relevant information rises as a result of the daily increase in the volume of data on the Internet.

**Manuscript received on 20 September 2023 | Revised Manuscript received on 26 September 2023 | Manuscript Accepted on 15 November 2023 | Manuscript published on 30 November 2023.**
* Correspondence Author (s)

**Aahana Saklecha***, Student, Department of Electrical Engineering, Shri Govindram Seksaria Institute of Technology and Science, Indore, Madhya Pradesh, India. E-mail: aahanajain2002@gmail.com, ORCID ID: 0009-0008-6433-1753

**Pragya Uplavdiya**, Student, Department of Information Technology, Shri Govindram Seksaria Institute of Technology and Science, Indore, Madhya Pradesh, India. E-mail: pragya.uplavdiya09@gmail.com, ORCID ID: 0009-0004-0253-9250

**Prof. M.P.S. Chawla**, Associate Professor, Department of Electrical Engineering, Shri Govindram Seksaria Institute of Technology and Science, Indore, Madhya Pradesh, India. E-mail: mpschawla@gmail.com



Systems for text summarizing can speedily automate the process of creating a summary from a lengthy text. Historically, researchers have focused on creating statistical models to accomplish this. Recently, various machine learning algorithms that can automatically createmodels have come under thorough investigations. A summary is a condensed, easily accessible text that includes the key points from one or more papers. A single document [12] or a collection of documents are used to create summaries. Multi-document summarizing is the creation of a summary from multiple documents. A 'query-biased' summary [13] highlights the user about queries. The drafting of topics and the inclusion of the most useful phrases are two aspects of topic summarization. The current work focuses on improving summaries by restricting the material to just the information that the user needs [14]. Summaries produced by humans are more expensive, and those produced by machines are often subpar. Researchers necessitate a number of efforts to produceaccurate, illuminating summaries [15].

## II. LITERATURE SURVEY

In natural language processing (NLP), text summarizing is a critical activity that seeks to automaticallyproduce a reduced version of a text, while maintaining its essential information. Recent years have seenconsiderable progress in the field of text summarization due to the continuous increase of digital content and the requirement for effective information processing and retrieval. In this review of the literature, athorough overview of the most recent approaches to text summarization using NLP has been discussed.

### A. Extractive Summarization

The most popular technique for text summarization is called extractive summarization, which is selecting the key words or sentences from the original text and putting them together into a summary. The Text Rank method, developed by Mihalcea and Tarau in 2004, is one of the earliest efforts in this field. It uses graph-based ranking to find the most crucial sentences in a document. Later, a number of additional approaches, including LSA (Steinberger et al., 2004) [6], Sum Basic (Vanderwende, 2007) [7], and Lex Rank (Erkan and Radev, 2004) [8], have been presented for extractive summarization. On a variety of datasets and domains, these strategies have yielded encouraging results [2].

### B. Abstractive Summarization

The more difficult method of summarization is creating a summary that is not restricted to the words or phrases from the original text. Instead, it seeks to produce a summary that conveys the text's main ideas in a manner that seems natural to the reader.

The RNN-based Encoder-Decoder model by Rush et al. (2015) [3], one of the pioneering studies in this field, creates a summary from the input content using a neural network architecture. Later, a number of additional approaches have been put out for abstractive summarization, including Pointer-Generator Networks (See et al., 2017) [4], Transformer- based models (Vaswani et al., 2017) [5], and GPT-based models (Radford et al., 2019) [5]. Although these techniques have produced encouraging results, they still need to be improved.

### C. Evaluation Metrics

It might be difficult to assess a summary's quality because there are so many factors to take into account, including relevance and coherence. For text summarization, a number of evaluation metrics have been put forth, including METEOR (Banerjee and Lavie, 2005) [9], BLEU (Papineni et al., 2002) [10], and ROUGE (Lin, 2004) [11]. These metrics are based on calculating the overlap between the generated summary and one or more reference summaries. These metrics do have some drawbacks, though, such as the fact that they don't take the summary's grammar and flow into account.

### D. Multi document Summarization

Summarizing numerous documents on a single subject is a more delicate task known as multi- document summarization. Finding the essential details and connections between colorful documents makes this work more delicate than summarizing a single text. For multi-document summarization, a number of approaches have been put forth, including Cluster Rank (Erkan and Radev, 2004) [12], MMR (Carbonell and Goldstein, 1998) [13], and MEAD (Radev etal., 2004) [14]. These methodologies give a summary from several documents using varied methods like clustering, sentencefusion, and graph- based ranking.

### E. Domain specific summarization

The task of summarizing text from a particular domain, similar as scientific publications, news stories, or legal documents, is known as domain-specific summary. This task necessitates the employment of technical models and procedures that can directly capture the domain's unique language and jotting style [23].

## III. MAIN STEPS FOR TEXT SUMMARIZATION

There are three main ways for summarizing documents. These are:

1. **Topic Identification:** The most prominent information in the text is identified. There are different ways for topic identification that are used which are Position, Cue Expressions, and word frequencies. Approaches are grounded on the position of expressions are the most useful styles for topic identification.
2. **Interpretation:** Abstract summaries need to go through an interpretation step. In this step, different subjects are fused in order to form a general content.
3. **Summary Generation:** In this step, the system uses a text generation system with acompression rate of 5-30% to give an acceptable summary quality [18].

## IV. MATERIALS AND METHODS

The majority of the current automatic text summarizing systems produce a summary using anextraction strategy. To produce extractive summaries, sentence extraction algorithms are frequently utilized. Assigning a numerical value to a sentence's summary is one way to come up with applicable sentences [1]. Based on the compression rate, choose the best sentences toproduce the document summary, a process known as sentence scoring. The compression rate is a pivotal element of the extraction procedure that's used to determine the proportion betweenthe length of the summary and the original text. The summary will grow in size and contain more irrelevant content as the compression rate increases [18]. While the summary becomes shorter due to its reduced compression rate and hence further information is lost.

## V. EXTRACTIVE SUMMARIZATION METHODS

### A. Term Frequency-Inverse Document Frequency (TF-IDF) method:

This numerical statistic, known as TF-IDF (Term Frequency-Inverse Document Frequency), gauges the significance of a word within a particular document. The TF-IDF value increases proportionally with the word's frequency in the document. This methodology primarily operates on the weighted term frequency and inverse sentence frequency framework. Sentence frequency refers to how many sentences in the document contain the term. These sentence vectors are further assessed for similarity to the query, and the highest-scoring ones are selected for inclusion in the summary. Summarization here is tailored to the specific query, making certain assumptions. Under this approach, if a sentence contains more specific words, it is deemed relatively more important. Typically, the focus is on nouns as target words, and the system conducts a comparison of term frequencies (TF) within the document. In this context, each sentence is treated as a separate document, and the document frequency (DF), which indicates how often the word appears in the entire collection of documents, is considered. The TF/IDF score is computed as follows [18]:

$$\frac{TF}{ID(w)} = DN\left(\frac{\log(1+tf)}{\log(df)}\right) \quad (1)$$

### B. Graph Theoretic approach:

This approach involves the creation of a node for each sentence present in a document. Nodes are linked together if the sentences they represent share common words, indicating a similarity above a predefined threshold. This representation has two primary outcomes. Firstly, it results in partitions within the graph, which are essentially distinct motifs found within the document and disconnected from other sub-graphs. Secondly, this graph-based approach can identify important sentences within the document. Sentences witha high level of connectivity (i.e., having a significant number of edges connecting to them) are considered important within their respective partitions, making them more likely candidates for inclusion in the summary. Referencing Figure No. 1 [18], we can observe an illustrative graph of a document.

It's apparent that the document encompasses approximately 3-4 subjects, and the encircled nodes can be recognized as instructive sentences because they share information with many other sentences in the document. Additionally, this graph-theoretic system can be easily adapted to assess both inter and intra-document similarity.

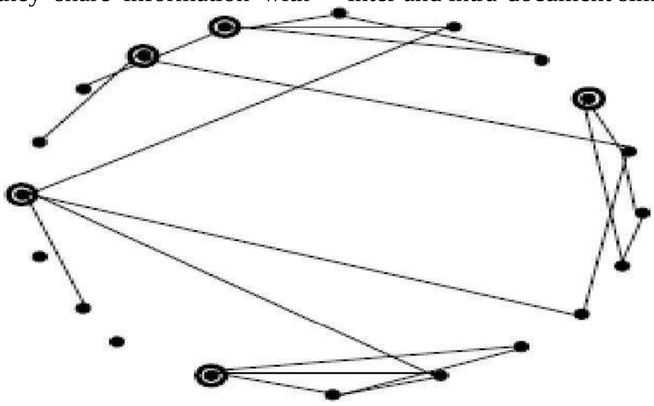

**Fig. 1: Graph Theoretic Approach [18]**

### C. Machine learning approach:

In this system, the training dataset serves as a point of reference, and the summarization process isconceptualized as a classification task. Sentences are categorized as either summary sentences or non- summary sentences based on the specific features they exhibit. The statistical learning of classification probabilities is accomplished using Bayes' rule [18],

$$P\ (s \in < \ S \mid F1, F2, \ldots, FN) = P(F1, F2, \ldots, FN \mid s \in \ S) * \frac{P(s \in S)}{P(F1, F2, \ldots, FN\ )} \quad (2)$$

where 's' represents a sentence from the document collection, 'F1', 'F2', ... 'FN' denote the features utilized in the classification process. 'S' represents the summary to be generated, and 'P(s∈< S | F1, F2, ..., FN)' signifies the probability that a sentence 's' will be selected to compose the summary, given that it possesses features 'F1', 'F2', ... 'FN'.

## VI. EVALUATING THE SUMMARIZATION SYSTEMS

Evaluation methodologies are helpful in assessing the summary's utility and reliability. It can be laborious to judge traits like readability, coherence, and comprehensibility. Experts estimate the quality of a summary using a manually operated expert system [19]. The measurement of the system'squality is carried out by counting the number of sentences chosen by the system that corresponds tothe ideal natural sentence.

| Case Study | Challenging Issues | Remedial Measures | Outcomes/ Advantages | Overall SuccessRate |
|---|---|---|---|---|
| Automatic text summarization of konkani texts using pre-trained word embeddings and deep learning. [15] | The size of the training dataset increased since a different dataset was required to learn word embedding. | This requisite was satisfied by applying fast Text because it was pre-trained and downloadable. Since fast Text was trained on Konkani documents from Common Crawl and Wikipedia, this was made practicable. | The final product closely matched the summaries created by the linguists. | According to the study, supervised deep multi-layer perceptron (MLP) methodologies can produce summaries of more advanced quality than unsupervised techniques. |
| Automatic Text Summarization using a MachineLearning Approach [1] | 1. There are not numerous illustrations ofhow to choose the applicablefeatures whenusing heuristics. 2. The production andanalysis of a text's entire rhetorical structure wouldbe unachievable with present text processing technology. | 1. It makes use ofdirectly andautomatically extricate statisticaland verbal elementsfrom the originaltext. 2. An agglomerative clustering algorithm is performed to thetext to produce thisapproximation ofthe structure. | According to the findings, scientistsshould center their works on the studyof further complex classifiers designed for the text- summarization task, or at the leastcompare and choose the top classifier from the traditionalones that are alreadyaccessible. | For all of the strategies, the accuracy andrecall ranks are noticeably lesser. Using Naive Bayes, the trainable summarizer produced the topmost results. |

Retrieval Number:100.1/ijsp.D1016113423
DOI:10.54105/ijsp.D1016.113423
Journal Website: www.ijsp.latticescipub.com



Published By:
Lattice Science Publication (LSP)

Figure 2 shows the architecture of the proposed summary refinement model.

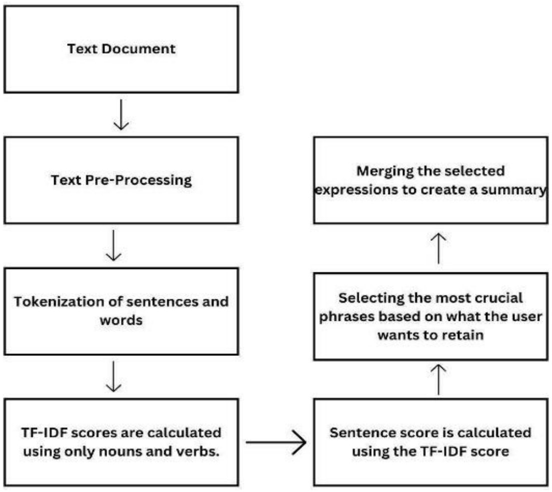


**Fig. 2: Text Summarization Data Flow Diagram [16]**

i. **Input:**
   - The process starts with a text document as input, which is the content that needs to besummarized.

ii. **Text Pre-Processing:**
   - Special characters are removed.
   - Text is converted to lowercase.
   - Punctuation is handled to prepare the text for analysis.

iii. **Tokenization:**
   - The text is divided into individual sentences and words.
   - Tokenization is necessary for further analysis.

iv. **TF-IDF Scores Calculation Using Nouns and Verbs:**
   - TF-IDF scores are calculated for each word or term in the document.
   - Only nouns and verbs are considered for this calculation.
   - This step helps identify the most important nouns and verbs in the text.

v. **Sentence Score Calculation Using TF-IDF:**
   - Each sentence in the text is assigned a score based on the TF-IDF scores of the nouns and verbs it contains.
   - Sentences with higher scores are considered more significant.

vi. **User Interaction for Phrase Selection:**
   - The user interacts with the system to specify which phrases or information they want toretain.
   - This could involve manual selection or setting criteria for automated selection.

vii. **Selecting Most Crucial Phrases:**
   - The system selects the phrases or sentences based on the user's preferences.
   - It retains only the most crucial information as specified by the user.

viii. **Merging Selected Expressions to Create a Summary:**
   - The selected phrases or sentences are merged together.
   - This merging process generates a concise summary of the text.
   - The summary captures the main points of the document as per the user's preferences.

ix. **Output:**
   - The final output is the summarized version of the original text, containing only theinformation the user wants to retain.

This flowchart outlines the steps involved in processing a text document and generating a summary, with a focus on extracting and retaining the most relevant information.[16]

Example System: **SUMMARIST (Hovy and Lin, 1998) [18]:**

There are three stages required for the systematic procedure as discussed below:

$$Summary = Subject\ Id + Interpretation + Generation$$

1. **Subject Identification Modules**: Positional significance, Cue Expressions (under construction), Word Counts, Discourse Structure (under construction).
2. **Subject Interpretation Modules**: Concept Counting/ Wave front, Concept Signatures (beingextended)
3. **Summary Generation Modules** (not yet constructed) Keywords, Template Gen, and Sent. Planner & Realizer. [17]

In the modules mentioned above, various techniques are employed to identify the subject of interest in the text. This includes assessing positional significance, cue expressions (though this module is under construction), counting words, and analyzing discourse structure. Subject Interpretation Module focuses on interpreting the identified subject. It employs methods like concept counting and wave front analysis, along with concept signatures which are being extendedto provide a deeper understanding of the subject. Although not fully constructed at the time of the system's description, the last stage is crucial for generating the actual summary. It includes sub-modules like keyword extraction, template generation, and sentence planning and realization, all of which contribute to creating a concise and coherent summary of the input text. Overall, SUMMARIST aims to systematically identify, interpret, and ultimately generate summaries from text, with an emphasis on subject understanding and effective summary creation. The system mentioned, SUMMARIST (Hovy and Lin, 1998), involves a systematic procedure with three key stages:

## VII. CONCLUSION

A vital aspect of text summarization is summary evaluation. Summaries can generally be assessed usingeither natural or foreign measures. Foreign techniques estimate summary quality through a task-based performance measure, similar as an information recovery- orientated task, while natural strategies aim to do so through assessment by humans. The important thing that's considered intriguing from the review that has been done in the results of the analysis which states that extractive summaries are fairly easier than abstractive summaries which are veritably complex and on the other hand, extractive summaries are still the subject of current favorite trends [20] as there are still numerous things that are a challenge for researcher scholars to do.

*Retrieval Number:100.1/ijsp.D1016113423*
*DOI:10.54105/ijsp.D1016.113423*
*Journal Website: www.ijsp.latticescipub.com*



*Published By:*
*Lattice Science Publication (LSP)*

It also can be seen that the most important features to produce a good summary are keywords, frequency, similarity, sentence position, sentence length, and semantics.In the future, these systems can be optimized to produce further accurate summaries. The future also holds scope for advancement in text summarization of documents in native language.

## DECALARION STATEMENT

| | |
|---|---|
| Funding/ Grants/ Financial Support | No, I did not receive. |
| Conflicts of Interest/ Competing Interests | No conflicts of interest to the best of our knowledge. |
| Ethical Approval and Consent to Participate | No, the article does not require ethical approval and consent to participate with evidence. |
| Availability of Data and Material/ Data Access Statement | Not relevant. |
| Authors Contributions | All authors have equal participation in this article. |

## REFERENCES


1. Joel Larocca Neto Alex A. Freitas Celso A. A. Kaestner In Automatic Text Summarization using aMachine Learning Approach.
2. https://www.analyticsvidhya.com/blog/2023/03/exploring-the-extractive-method-of-text-summarization/
3. Suzuki, Jun & Nagata, Masaaki. (2016). RNN-based Encoder-decoder Approach with WordFrequency Estimation.
4. Abigail See, Peter J. Liu, Christopher D. Manning In Get To The Point: Summarization withPointer-Generator Networks
5. Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia In Attention is all you need
6. Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understandingby generative pre-training.
7. Steinberger, Josef & Jezek, Karel In Using Latent Semantic Analysis in Text Summarization andSummary Evaluation.
8. Lucy Vanderwende, Hisami Suzuki, Chris Brockett, Ani Nenkova In Beyond SumBasic: Task- focused summarization with sentence simplification and lexical expansion
9. Gunes Erkan, Dragomir R. Radev In LexRank: Graph-based Lexical Centrality as Salience in TextSummarization
10. Satanjeev Banerjee and Alon Lavie In METEOR: An Automatic Metric for MT Evaluation withImproved Correlation with Human Judgments
11. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu In Bleu: a Method forAutomatic Evaluation of Machine Translation
12. Chin-Yew Lin In ROUGE: A Package for Automatic Evaluation of Summaries
13. X. Wan and J. Xiao (2010), ―Exploiting neighborhood knowledge for single document summarization and key phrase extraction‖, ACM Trans. Inf. Syst., vol. 28, pp. 8:1–8:34. https://doi.org/10.1145/1740592.1740596
14. J. G. Conrad, J. L. Leidner, F. Schilder, and R. Kondadadi (2009), ―Query-based opinion summarization for legal blog entries‖, Proceedings of the 12th International Conference on Artificial Intelligence and Law, pp. 167–176. https://doi.org/10.1145/1568234.1568253
15. G. Amati and C. J. Van Rijsbergen (2002), ―Probabilistic models of information retrieval based on measuring the divergence from randomness, ‖ ACM Trans. Inf. Syst., vol. 20, pp. 357– 389. https://doi.org/10.1145/582415.582416
16. A Redundancy Elimination Approach towards Summary Refinement Esther Hannah. M [a], Saswati Mukherjee [b], Sakthi Balaramar [c]
17. Aakash Srivastava; Himanshu Daharwal ; Kamal Chauhan ; Nikhil Mukati ; Pranoti Shrikant Kavimandan In Text Summarizer Using NLP (Natural Language Processing), Volume 6 Issue 1 2022, pp. 214.
18. Eduard Hovy and Chin-Yew Lin. 1998. Automated Text Summarization and theSummarist System, pp. 2.
19. Eduard Hovy and Daniel Marcu in Automated Text summarization Tutorial — COLING/ACL ‘98.
20. Mr. S. A. Babar, M. Tech-CSE, RIT In Text Summarization: An Overview, pp.2-4.
21. Esther Hannah. M, Saswati Mukherjee, Sakhti Balaramar in A redundancy elimination approach towards summary refinement, pp. 249-250
22. Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi in Review of automatic textsummarization techniques & methods.
23. Varun Pandya In Automatic Text Summarization of Legal Cases: A Hybrid Approach.
24. K. A. R. Issam, S. Patel*, and S. C. N., “Topic Modeling Based Extractive Text Summarization,” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 6. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 1710–1719, Apr. 30, 2020. doi: 10.35940/ijitee.f4611.049620. Available: http://dx.doi.org/10.35940/ijitee.F4611.049620
25. V. Kanaparthi, “Examining Natural Language Processing Techniques in the Education and Healthcare Fields,” International Journal of Engineering and Advanced Technology, vol. 12, no. 2. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 8–18, Dec. 30, 2022. doi: 10.35940/ijeat.b3861.1212222. Available: http://dx.doi.org/10.35940/ijeat.B3861.1212222
26. S. K* et al., “Youtube Video Ranking: A NLP based System,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 4. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 1370–1375, Nov. 30, 2019. doi: 10.35940/ijrte.d7303.118419. Available: http://dx.doi.org/10.35940/ijrte.D7303.118419
27. N. N.S and S. A, “Malware Detection using Deep Learning Methods,” International Journal of Innovative Science and Modern Engineering, vol. 6, no. 6. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 6–9, Apr. 15, 2020. doi: 10.35940/ijisme.f1218.046620. Available: http://dx.doi.org/10.35940/ijisme.F1218.046620
28. Dr. D. Devrapalli et al., “Effective Text Processing utilizing NLP,” Indian Journal of Artificial Intelligence and Neural Networking, vol. 2, no. 1. Lattice Science Publication (LSP), pp. 1–7, Dec. 30, 2021. doi: 10.54105/ijainn.b3873.122121. Available: http://dx.doi.org/10.54105/ijainn.B3873.122121


## AUTHOR PROFILE

**Aahana Saklecha,** a proficient student pursuing B.Tech. in Electrical Engineering from SGSITS, Indore with a strong academic background, having completed her 12th and 10th grades with outstanding scores. She pursued Data Science and Machine Learning training from IIT Madras and excelled in foundational courses. She has practical experience as an engineering intern, focusing on Electric Vehicle techno-commercial analysis. She has undertaken academic projects in speed control of DC motors and home automation, displaying her proficiency in microcontrollers and IoT. Her technical skills encompass Python, C++, data analytics, and web development, and she has contributed to the development of the Entrepreneurship Cell website. Her leadership roles include Head of External Affairs at E-Cell, where she currently is the President.

**Pragya Uplavdiya,** dedicated and accomplished Data Science and Programming enthusiast with a strong academic background and a passion for leveraging technology to solve real-world challenges. Her diverse skill set spans languages like C/C++, Python, and R, and Web designing skills including HTML, CSS and Javascript. She is currently pursuing a B.Sc. in Data Science and Programming at the Indian Institute of Technology (IIT), Madras, and a B.Tech. in Information Technology at Shri G.S. Institute of Technology and Science, Indore. In her role as Training and Placement Coordinator at SGSITS, Indore and facilitated campus drives and built strong industry partnerships, aiming for continuous growth. She has also contributed to the aerospace community as a member of Team YAN.

**Prof. M.P.S. Chawla,** Ex-Professor-Incharge (head), library, Associate Professor in Electrical Engineering Department, SGSITS, Indore-452003 (M.P.)-India, immediate past chairman, 2017- 2018, IEEE M.P. sub-section. He received gold medals in B.E.(Electrical) and M.E. (Power Electronics) degree SGSITS, Indore, India from Electrical Engineering Department in 1988 and 1992, respectively. He was appointed as" Associated Editor in Chief Chair", on 14th december2018 in Blue Eyes Intelligence Engineering and Sciences Publications, India. His special research interests are in power electronics, devices, intelligence instrumentation, biomedical engineering, signal processing, advanced instrumentation systems, soft computing, higher order statistical techniques and control systems. He was also the Ex-PG Coordinator of SGSITS Indore.