International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

AN OUTLIER DETECTION-DRIVEN FRAMEWORK FOR ENHANCED INFORMATION RETRIEVAL

Wahid Ali[1], Mohd Waris Khan[2*]

[1,2]Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India

wahidali@student.iul.ac.in, wariskhan070@gmail.com

Abstract

The growing complexity and volume of data have made information retrieval (IR) systems more challenging, requiring innovative techniques for enhanced accuracy and efficiency. This paper proposes a novel framework for information retrieval, which integrates outlier detection with machine learning (ML) algorithms to build a robust feedback model aimed at refining the retrieval process. The framework is designed to automatically identify and handle outliers data points that deviate significantly from the norm by leveraging outlier detection techniques such as statistical methods, density-based models, and distance-based approaches. The core of this framework lies in its ability to adaptively improve the performance of IR systems over time through continuous feedback loops. Using ML models such as supervised learning for training the retrieval system based on labelled data and unsupervised learning to automatically detect anomalies or outliers. This framework can dynamically adjust its retrieval processes to avoid irrelevant or erroneous data that might distort search results. A key feature of the proposed model is its ability to integrate both active and passive feedback mechanisms. Active feedback involves the system's real-time adaptation to user input, such as clicks, relevance ratings, or query modifications. Through the deployment of this outlier detection-driven framework, we aim to enhance the accuracy and user satisfaction of information retrieval systems, making them more resilient to noise, errors, and irrelevant data. The approach ensures that the system not only provides the most relevant information but also adapts effectively to evolving user behaviours and data trends.

Index terms: Information Retrieval (IR), Outlier Detection, Machine Learning (ML), Feedback Mechanism, Supervised Learning, Unsupervised Learning

Introduction

The rapid growth of data across various domains has significantly influenced the performance of traditional Information Retrieval (IR) systems. As these systems become increasingly complex and handle diverse data types, the need for more accurate and efficient retrieval methods has become paramount. Information retrieval systems typically rely on ranking and relevance-based algorithms to provide users with the most pertinent results [1, 2]. However, these systems are often susceptible to errors caused by irrelevant, redundant, or misleading data. Outlier's data points that deviate significantly from the general distribution of information pose a particular challenge, as they can skew search results, degrade system performance, and

Received: August 01 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

confuse users [3]. To address this challenge, we propose an innovative framework that integrates outlier detection techniques with machine learning (ML) to enhance the accuracy and robustness of information retrieval systems [4, 5]. Outlier detection is critical in identifying and mitigating the impact of anomalous data, which may include erroneous queries, irrelevant documents, or noise. By incorporating supervised and unsupervised machine learning models, our framework enables the system to continuously learn from both explicit user feedback (such as clicks or relevance ratings) and implicit signals (such as browsing patterns or dwell time).

The key advantage of this framework lies in its adaptive feedback mechanism, which not only refines search results but also learns to identify evolving patterns in user behavior and data trends. By integrating outlier detection into the feedback loop, the system can dynamically filter out irrelevant data and focus on delivering the most relevant and accurate results. This feedback-driven approach ensures that the system continuously improves, enhancing its predictive capabilities and providing users with more personalized and contextually accurate information over time. The ability to detect and correct for outliers in real-time makes this framework highly effective in real-world applications, where data is noisy, dynamic, and constantly evolving. By bridging the gap between anomaly detection and information retrieval, this framework not only promises to improve the quality of search results but also offers a more scalable and efficient solution to information retrieval in the face of increasing data complexity.

Information Retrieval (IR) Systems

Information Retrieval (IR) systems are designed to retrieve relevant information from large collections of data based on user queries. These systems play a crucial role in accessing and organizing information, and are widely used in various applications such as web search engines, digital libraries, e-commerce platforms, and academic databases. The primary goal of IR systems is to deliver documents or data that are most relevant to the user's information needs, based on the query provided. At the core of traditional IR systems lies the concept of document ranking, where documents are scored according to their relevance to a given search query, often using algorithms like TF-IDF (Term Frequency-Inverse Document Frequency), Vector Space Model, or more advanced probabilistic models [6, 7]. These systems typically leverage metadata, keywords, or full-text indexing to organize large amounts of unstructured data and make it accessible. The performance of IR systems has traditionally been assessed in terms of precision and recall, with the goal of minimizing irrelevant results while maximizing the retrieval of relevant documents. However, as the volume of data increases, the complexity of providing relevant and accurate search results has grown. This is further compounded by challenges such as ambiguity in queries, synonymy, polysemy, and data noise. With the advent of more sophisticated machine learning techniques and the expansion of unstructured data sources, IR systems have evolved to include semantic search, personalized search, and context-aware retrieval to improve the quality and relevance of search results [8, 9]. Despite these advancements, traditional IR systems often struggle with issues such as handling ambiguous queries, filtering out irrelevant results, and adapting to changing user preferences over time. As data becomes increasingly varied and complex, IR systems must incorporate more adaptive and intelligent mechanisms to refine their search capabilities, making them more efficient and

Received: August 01 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

user-friendly [10, 12]. This necessitates the incorporation of techniques such as outlier detection, anomaly detection, and machine learning, which can provide deeper insights into the quality of search results and improve the overall user experience by filtering out noise and irrelevant data [11, 12]. In figure 1, illustrating Information Retrieval (IR) Systems. It highlights key aspects like Search Engines, Components, Techniques, Challenges, and Applications, with further details on processes such as Query Processing, Ranking, and Indexing, as well as techniques like Machine Learning and Natural Language Processing.

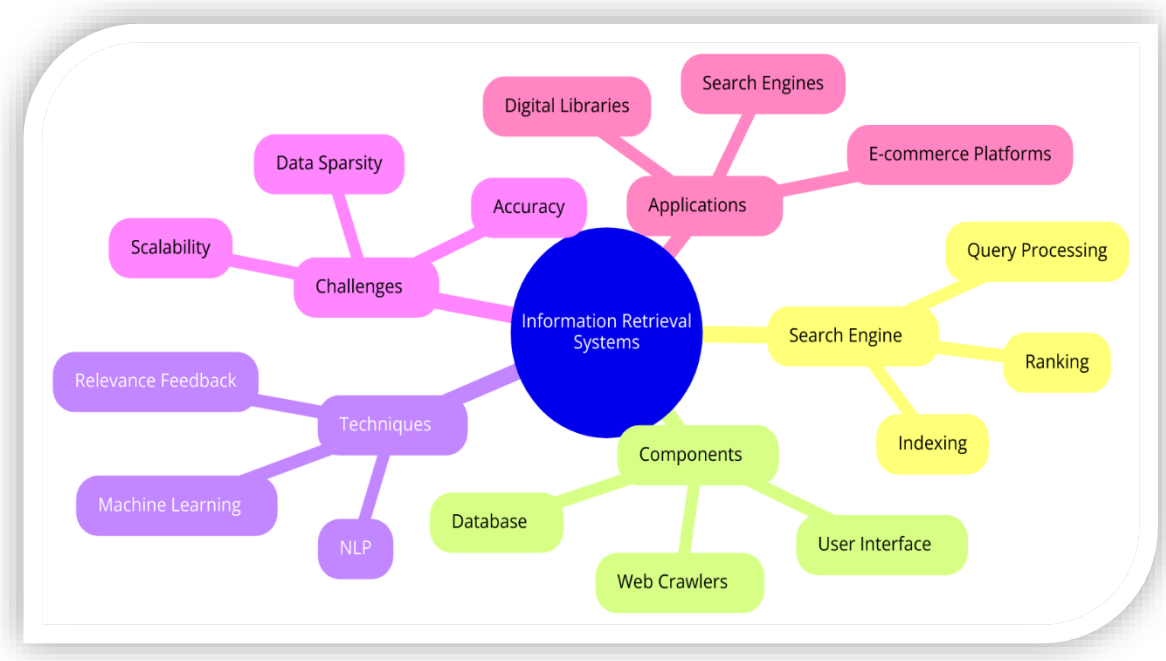

Fig 1: Information Retrieval (IR) Systems

The Framework

“An Outlier Detection-Driven Framework for Enhanced Information Retrieval" is a framework designed to improve the performance of Information Retrieval (IR) systems by using outlier detection techniques. This means that it will identify and handle outliers (data points that are significantly different from the usual data patterns) and use this information to make the IR system more effective and accurate.

Outlier Detection: The framework will first identify outliers in the data, which could be queries or documents that do not fit the typical patterns. These outliers are often ignored or poorly handled by traditional IR systems.

Improved Query Handling: When a user submits a query that deviates from the usual patterns (an outlier query), the framework will better understand and process it, providing more relevant search results that traditional systems might miss.

Anomalous Document Retrieval: Outliers can also appear in documents. If a document is extremely relevant but doesn’t fit the typical patterns, a traditional IR system might overlook

Received: August 01 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

it. The framework will highlight such documents, improving the results by considering those outliers.

Data Quality Improvement: By detecting and handling outliers, the framework can improve the overall quality of the data used by the IR system. It can remove irrelevant or misleading data, which in turn improves the accuracy and precision of the search results.

Handling Noisy Data: Data in IR systems is often noisy, meaning it contains irrelevant or incorrect information. This framework will identify and filter out such noisy data, resulting in more accurate and relevant search results.

This framework transforms traditional IR systems into more adaptive and intelligent systems by focusing on detecting and handling outliers. By doing so, it improves relevance, accuracy, and precision in search results, especially when the data contains unusual or abnormal patterns that traditional IR methods might not handle well.

In figure 2, framework outlines a comprehensive approach for enhancing an Information Retrieval (IR) system through outlier detection. It involves data preparation, model training, optimization, evaluation, and continuous recalibration based on user feedback. The key processes aim to improve the system's ability to handle unusual or abnormal data (outliers), thereby improving the relevance, accuracy, and quality of the search results. The framework is organized into four stages, each focusing on a crucial step in building and optimizing an Outlier Detection-Driven Information Retrieval (IR) System.

Stage 1: Data Preparation

This stage focuses on gathering and preparing the data before it's used for building and training the model. It's essential because the quality of data directly affects the model's performance. The following steps are involved in this stage:

- **Input Data**:

o The first step is gathering the raw data, which could be queries (search input) or documents (the content to be searched). This data will later be processed and used by the system.

- **Outlier Detection Process**:

o **Outliers** are data points that differ significantly from the rest of the data. In IR systems, these could be queries or documents that are too different from others. Outlier detection identifies these points, which may either need to be handled separately or removed from the dataset if they are irrelevant or noisy.

- **Data Pre-Processing**:

o After detecting potential outliers, the data needs to be processed so that it's in a format that the model can understand. This includes several sub-processes:

Received: August 01 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Feature Extraction**: In this step, the most important attributes (or features) of the data are extracted. For example, in an IR system, features could be terms from documents or keywords in queries that are crucial for searching.

- **Data Cleaning**: This step involves cleaning the data to remove errors, missing values, duplicates, or irrelevant information. It ensures the data is accurate and reliable.

- **Data Transformation:** Sometimes, the data may need to be transformed or restructured into a different format (e.g., converting text to numerical data) so that the model can use it.

- **Data Normalization:** This step standardizes the data so that different features have a comparable scale. For example, the range of values in the dataset might be adjusted to fall between 0 and 1, which helps models like clustering or machine learning algorithms to work effectively.

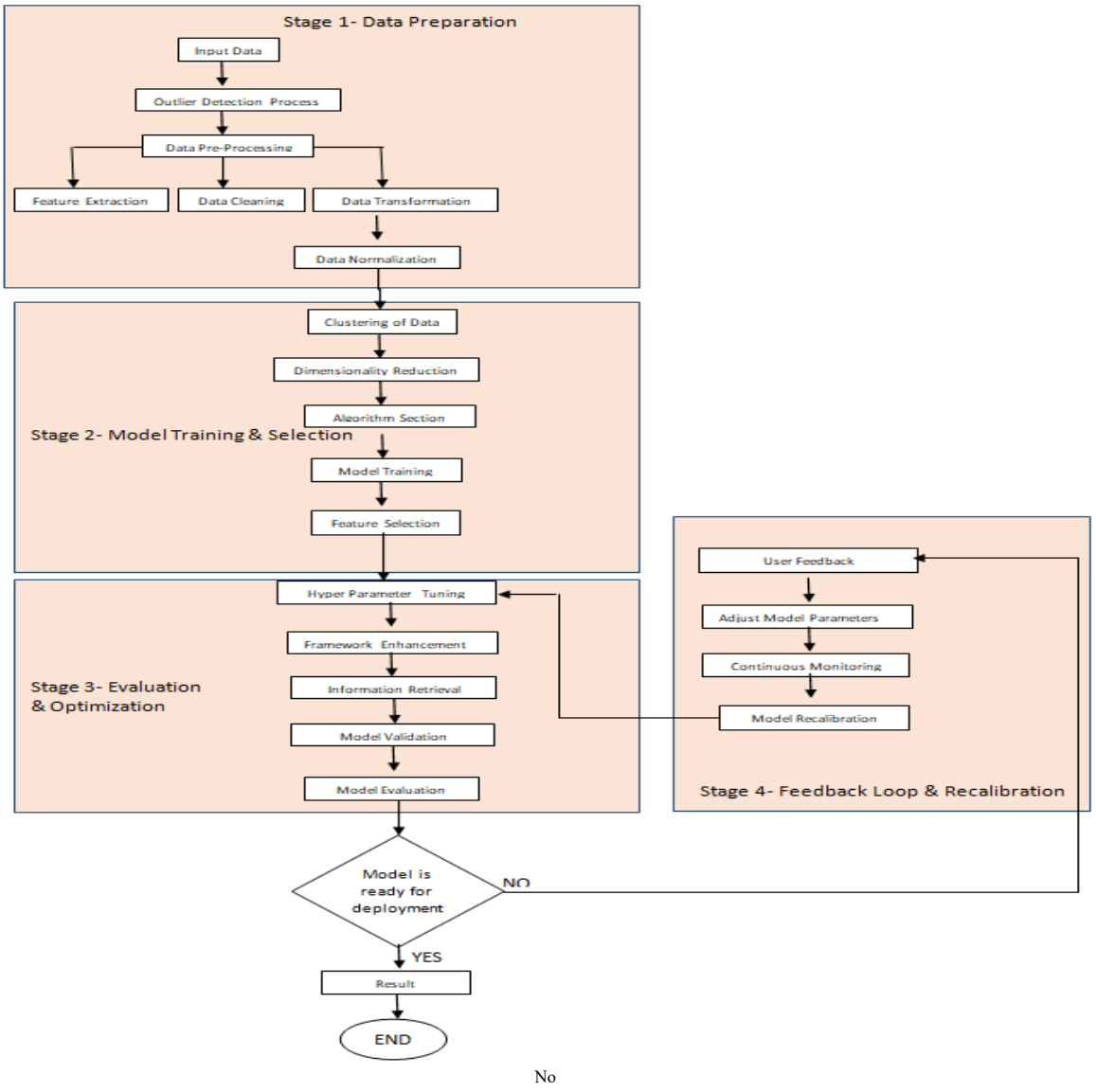

Fig 2: Proposed Framework

Received: August 01 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Stage 2: Model Training & Selection

Once the data is prepared, the next stage is about choosing the right algorithm, training the model, and selecting the best features. This stage is crucial for developing a system that can identify outliers and improve search results.

- **Clustering of Data:**

o Clustering is the process of grouping similar data points together. This can help the system understand the relationships between different queries and documents. By grouping data into clusters, the model can identify outliers as points that do not belong to any of the clusters.

- **Dimensionality Reduction:**

o In many datasets, there are too many features, which can make the model complex and slow. Dimensionality reduction reduces the number of features, keeping only the most essential ones that carry the most information. Techniques like Principal Component Analysis (PCA) are used to reduce complexity without losing important patterns.

- **Algorithm Selection:**

o Here, the most suitable machine learning or outlier detection algorithm is chosen. Depending on the data type and problem, the model could use methods like k-Nearest Neighbors (k-NN), Decision Trees, or more advanced techniques like Support Vector Machines (SVM) or Neural Networks. The right algorithm will help in detecting outliers effectively.

- **Model Training:**

o In this step, the selected algorithm is trained on the prepared data. The model learns to recognize patterns, relationships, and outliers from the data. Training involves adjusting the internal parameters of the model so it can make predictions or classifications.

- **Feature Selection:**

o After training, the model is evaluated to determine which features are the most important. Feature selection helps to improve the model's efficiency and accuracy by removing irrelevant or redundant features that do not contribute much to the prediction.

Stage 3: Evaluation & Optimization

After training the model, it's time to evaluate how well it performs, optimize its parameters, and ensure it can be used in an IR system for better results. This stage involves:

- **Hyperparameter Tuning:**

o Hyperparameters are settings that control the training process of a model (like learning rate, number of clusters, etc.). Tuning these hyperparameters optimizes the model's performance. Different combinations of hyperparameters are tested to find the best configuration.

Received: August 01 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Framework Enhancement:**

o Based on the evaluation of the model, this step focuses on improving the overall framework. If the model isn't performing as expected, this could involve revising the outlier detection process, reprocessing data, or even using a different algorithm.

- **Information Retrieval System:**

o In this step, the trained model is integrated into the actual IR system. The goal is to see how well the model performs in the real-world scenario of information retrieval. The system should be able to retrieve relevant documents and handle outlier queries more efficiently.

- **Model Validation:**

o Model validation involves testing the model on new, unseen data to check its generalizability. A model that performs well on training data but poorly on unseen data is overfitting. Validation helps to ensure that the model can handle different data scenarios.

- **Model Evaluation:**

o After validation, the model is evaluated using various performance metrics like accuracy, precision, recall, or F1-score. This evaluation will determine if the model is ready for deployment or if further optimization is required**.**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Stage 4: Feedback Loop & Recalibration

Once the model is deployed, it's crucial to continuously improve and adapt it based on real-world usage. The feedback loop helps ensure that the system remains effective over time, especially as new data comes in. The process is as follows:

- **User Feedback:**

o After the model is deployed, feedback is collected from users to understand how well the system is performing. If users find the search results less relevant or inaccurate, that feedback will be valuable for improvement.

- **Adjust Model Parameters:**

Received: August 01 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Based on the feedback, the model parameters might need to be adjusted. For example, the system might need to change how outliers are identified or tweak the feature selection to improve retrieval.

- **Continuous Monitoring:**

- The model is continuously monitored to detect any performance issues or changes in data patterns that may require attention. If performance drops or new outliers appear, the system must be updated to handle these changes.

- **Model Recalibration:**

- Based on continuous monitoring and feedback, the model is recalibrated regularly. This recalibration ensures the model remains accurate and efficient over time, adapting to new types of data or queries.

Discussion and Conclusion

The Outlier Detection-Driven Framework for Enhanced Information Retrieval introduces a novel approach to improving search accuracy by addressing anomalous queries and documents that are often overlooked by traditional IR systems. By integrating outlier detection into the IR pipeline, the framework enhances data quality, filters noise, and highlights unusual yet potentially relevant information. Its structured design, comprising data preparation, model training and selection, evaluation and optimization, and feedback with recalibration, ensures continuous adaptability to changing data patterns, making it more robust and intelligent compared to conventional methods. Despite its strengths, the framework faces certain challenges such as computational complexity, sensitivity to algorithm parameters, and the need for scalable implementations to handle large datasets efficiently. However, the inclusion of a feedback and recalibration loop provides resilience against data drift, ensuring long-term effectiveness. Overall, the framework offers a strong foundation for next-generation IR systems that can deliver more relevant, precise, and context-aware results in dynamic and noisy information environments. The integration of outlier detection during data preparation, model training, evaluation, and continuous feedback allows the system to adapt and evolve with user needs and changing data patterns. This approach not only improves the relevance and accuracy of retrieved information by filtering out noise and irrelevant data but also enhances the system's adaptability and efficiency, providing a more robust and user-centric information retrieval experience.

Acknowledgement

This work is acknowledged under Integral University manuscript no. IU/R&D/2025-MCN0003918.

References

1. Kanev, A.; Nasteka, A.; Bessonova, C.; Nevmerzhitsky, D.; Silaev, A.; Efremov, A.; Nikiforova, K. Anomaly Detection in Wireless Sensor Network of the "Smart Home"

Received: August 01 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

System. In Proceedings of the 2017 20th Conference of Open Innovations Association (FRUCT), St. Petersburg, Russia, 3–7 April 2017; pp. 118–124.

2. Kumar Dwivedi, R.; Pandey, S.; Kumar, R. A Study on Machine Learning Approaches for Outlier Detection in Wireless Sensor Network. In Proceedings of the 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 11-12 Jan. 2018; pp. 189–192, doi:10.1109/CONFLUENCE.2018.8442992.
3. Jones, P.J., James, M.K., Davies, M.J., Khunti, K., Catt, M., Yates, T., Rowlands, A.V., Mirkes, E.M., 2020. Filterk: A new outlier detection method for k-means clustering of physical activity. Journal of Biomedical Informatics 104, 103397.
4. Paja, W., Pancerz, K., Pekala, B., Sarzyński, J., 2021. Application of the fuzzy logic to evaluation and selection of attribute ranges in machine learning, in: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6.
5. Yahyaoui, A.; Abdellatif, T.; Attia, R. READ: Reliable Event and Anomaly Detection System in Wireless Sensor Networks. In Proceedings of the 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2018; pp.193–198, doi:10.1109/WETICE.2018.00044.
6. Stojanović, B., Božić, J., Hoffer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E., Runevic, J., 2021. Follow the trail: Machine learning for fraud detection in fintech applications. Sensors 21.
7. Alhakami, H., Alhakami, W., Baz, A., Faizan, M., Khan, M. W., & Agrawal, A. (2022). Evaluating intelligent methods for detecting COVID-19 fake news on social media platforms. *Electronics, 11*(15), 2417. https://doi.org/10.3390/electronics 11152417
8. Varish, N., Hasan, M. K., Khan, A., Zamani, A. T., Ayyasamy, V., & Islam, S. (2023). Content-based remote sensing image retrieval method using adaptive tetrolet transform based GLCM features. *Journal of Intelligent & Fuzzy Systems, 44*(6), 9627–9650. https://doi.org/10.3233/JIFS-231083.
9. Srivastava, S., & Ahmed, T. (2021). Feature-based image retrieval (FBIR) system for satellite image quality assessment using big data analytical technique. *Psychology and Education, 58*(2).
10. Ahamad, M. K., & Bharti, A. K. (2021). Validation of clustering-based framework using unsupervised machine learning. In *2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM)* (pp. 1–6). IEEE. https://doi.org/10.1109/SASM51857.2021.9841205
11. Shabbir, M., Kumar, R., Khan, M. W., & Singh, H. (2025). Empirical analysis of computationally intelligent technique for software risk prediction. *Journal of Cybersecurity and Information Management, 17*(2), 200–209.
12. Trinh, V.V.; Tran, K.P.; Mai, A.T. Anomaly Detection in Wireless Sensor Networks via Support Vector Data Description with Mahalanobis Kernels and Discriminative Adjustment. In Proceedings of the 4th NAFOSTED Conference on Information and Computer Science Anomaly, Hanoi, Vietnam, 24-25 Nov. 2017; pp. 567–586, https://doi:10.1007/978-1-4615-3792-2_30.

Received: August 01 2025