RESEARCH PAPER

A Hybrid Xgboost–Pso Framework for Detecting Redundant Drug Data Access in Pharmaceutical Information and Drug Delivery Systems

M. Sunil Kumar[1], P. Neelima[2], D. Geetha[3], Dr. D. Ganesh[4], Katheeja Khanam Pathan[5], Y. Yethish[6]

[1] Professor & Dean - P&M, Department of Computer Science and Engineering, School of Computing, Mohan Babu University (erstwhile Sree Vidyanikethan Engineering College), Tirupathi, AP, India. Email: sunilmalchi1@gmail.com

[2] Assistant Professor, Department of CSE, School of Engineering and Technology, SPMVV, Tirupathi, AP, India. Email: neelima.pannem@gmail.com

[3] Assistant Professor, Department of CSE, Vignan's Institute of Management and Technology for Women, Hyderabad, TS, India. Email: dgeetha@vmtw.in, geethakrish18@gmail.com

[4] Associate Professor, Department of CSE, School of Computing, Mohan Babu University (erstwhile Sree Vidyanikethan Engineering College), Tirupathi, AP, India. Email: dgani05@gmail.com

[5] Assistant Professor, Department of AIML, Faculty of Engineering and Technology, Jain University, Kanakapura Rd, Bengaluru, Karnataka, India. Email: aamnakhan521@gmail.com

[6] Department of CSE, School of Computing, Mohan Babu University (erstwhile Sree Vidyanikethan Engineering College (Autonomous)), Tirupati, AP, India. Email: yethish.2010@gmail.com

Received: 20th Feb, 2026 | **Revised:** 4th Mar, 2026 | **Accepted:** 25th Mar, 2026 | **Available Online:** 10th Apr, 2026

ABSTRACT

Efficient management of pharmaceutical data is critical for ensuring the integrity, security, and optimal utilization of drug-related information in modern healthcare and drug delivery systems. Redundant access to drug datasets, including repeated downloads of clinical, formulation, or patient-related data, can lead to inefficiencies, increased system load, and potential security risks. This study proposes a Data Access Duplication Alert System (DADAS) tailored for pharmaceutical informatics environments to detect and prevent redundant data access in real time.

The proposed system integrates Extreme Gradient Boosting (XGBoost) with Particle Swarm Optimization (PSO) to achieve high accuracy in identifying duplicate data access patterns. PSO is employed for optimal feature selection, while XGBoost performs classification based on user behavior, drug data access frequency, and metadata attributes. The system enables real-time monitoring and alert generation, thereby supporting secure and efficient drug data management.

Experimental results demonstrate that the proposed hybrid model outperforms traditional approaches, achieving improved accuracy, precision, and recall. The framework contributes to enhancing pharmaceutical data governance, optimizing drug delivery workflows, and supporting secure healthcare systems.

KEYWORDS

Pharmaceutical Data Management, Drug Delivery Systems, Data Duplication Detection, XGBoost, Particle Swarm Optimization, Healthcare Informatics, Machine Learning, Data Security

How to cite this article: Sunil Kumar M, Neelima P, Geetha D, Ganesh D, Pathan KK, Yethish Y. A Hybrid Xgboost–Pso Framework for Detecting Redundant Drug Data Access in Pharmaceutical Information and Drug Delivery Systems. Int J Drug Deliv Technol. 2026;16(29s):1015-1022. DOI: 10.25258/ijddt.16.29s.126

1. INTRODUCTION

The increasing adoption of digital technologies in pharmaceutical sciences and healthcare systems has

led to the generation and frequent access of large volumes of drug-related data, including clinical records, formulation data, and patient treatment information. Efficient management of such data is essential for ensuring the reliability and effectiveness of drug delivery systems. However, redundant access to pharmaceutical datasets such as repeated downloads of the same drug-related information can lead to inefficiencies, increased computational overhead, and potential security vulnerabilities [1][2].

Traditional data management systems primarily focus on access control and logging mechanisms, which are often insufficient for detecting repeated or redundant data access patterns in real time[3]. In pharmaceutical environments, such redundancies may affect system performance, compromise data integrity, and increase the risk of unauthorized data exposure.

To address these challenges, this study proposes an intelligent Data Access Duplication Alert System (DADAS) designed specifically for pharmaceutical informatics systems. The system leverages machine learning techniques to analyze user behavior, detect redundant data access, and generate real-time alerts. By integrating advanced predictive models, the system enhances data governance and supports efficient drug delivery workflows[4][5].

In the era of digital transformation, pharmaceutical and healthcare systems generate, manage, and access vast volumes of drug-related data, including clinical records, formulation details, and patient-specific treatment information. Ensuring the integrity, security, and optimal utilization of such data has become increasingly important as healthcare systems adopt data-driven technologies. One critical yet often overlooked issue in this context is the repeated access or retrieval of the same pharmaceutical dataset within a short time frame. This phenomenon, referred to as redundant drug data access, can lead to inefficiencies such as increased computational costs, unnecessary storage usage, excessive network bandwidth consumption, and potential exposure of sensitive healthcare information[6].

Conventional logging mechanisms and access control systems typically operate in a static and reactive manner, offering limited visibility into repeated access patterns. These systems are often inadequate in large-scale, distributed healthcare environments where multiple users—including clinicians, researchers, and pharmacists frequently access shared drug-related datasets. As a result, pharmaceutical systems face increased risks related to performance degradation, data redundancy, and potential violations of regulatory standards [7][8].

To address these challenges, this study proposes a Data Access Duplication Alert System (DADAS) specifically designed for pharmaceutical informatics environments. The system employs real-time monitoring, metadata analysis, and automated alert mechanisms to detect and mitigate redundant access to drug-related data. By collecting and analyzing key metadata attributes such as user identity, drug dataset identifiers, access timestamps, and frequency of access, the system evaluates duplication risk and generates timely alerts for both users and administrators. This facilitates responsible data usage while supporting compliance with healthcare regulations and data governance policies [9].

The proposed system is designed with a flexible, cloud-compatible architecture, enabling seamless integration into existing pharmaceutical and healthcare infrastructures. By incorporating machine learning techniques, including clustering and anomaly detection models, the system adapts to evolving access patterns, reduces false positives, and enhances detection accuracy over time. The framework supports efficient management of drug data workflows, thereby improving the performance and reliability of drug delivery systems [10].

Redundant access to pharmaceutical data is particularly common in environments such as hospitals, research institutions, and clinical laboratories, where multiple stakeholders frequently interact with shared datasets. In such scenarios, repeated access can overload system resources, disrupt real-time analytics, and increase the risk of unauthorized data exposure. Moreover, handling sensitive patient and drug-related data requires strict adherence to regulatory standards, making it essential to monitor and control access patterns effectively [11].

To overcome these limitations, the proposed DADAS framework introduces an intelligent middleware layer between the data access interface

and backend storage systems. This layer continuously monitors user activity, analyzes behavioral patterns, and identifies redundant access events with contextual awareness. Unlike traditional systems, it distinguishes between meaningful data access and unnecessary repetition, thereby optimizing resource utilization and improving user experience [12].

The system employs a hybrid machine learning approach to achieve high accuracy in duplication detection. Techniques such as K-Means clustering, DBSCAN, Autoencoders, and Isolation Forests are utilized to analyze patterns in pharmaceutical data access. These models are trained to differentiate between legitimate usage and potentially redundant or suspicious activities. Furthermore, reinforcement learning mechanisms enable the system to adapt dynamically based on user feedback and evolving data access trends. Through its integrated alerting mechanisms and user-friendly interface, the system empowers administrators to take proactive actions, such as restricting redundant access, providing cached data, or ensuring compliance with pharmaceutical data governance policies [13].

2. LITERATURE REVIEW

Recent advancements in pharmaceutical informatics have emphasized the importance of secure and efficient data management systems. Existing approaches primarily rely on access control mechanisms, encryption, and logging frameworks to ensure data security. However, these methods lack the ability to proactively detect redundant access patterns in drug-related datasets.

Machine learning techniques such as anomaly detection, clustering, and behavioral analytics have been widely applied in healthcare data analysis. Methods including Autoencoders, Isolation Forests, and ensemble learning models have shown effectiveness in identifying abnormal patterns. However, their application in detecting redundant pharmaceutical data access remains limited[14][15].

This study extends existing research by introducing a hybrid machine learning framework specifically designed to detect duplication in drug data access, thereby improving efficiency and security in pharmaceutical systems.

Intelligent data management solutions are becoming more and more necessary as a result of the growing volume and sensitivity of digital data in industries including healthcare, banking, and scientific research. The main goals of traditional systems are data security by logging, encryption, and access control. Recent research, however, shows that these approaches fall short in proactive ways to monitor duplicate data access, particularly when it comes to repeated downloads that are frequently overlooked in large-scale settings.

A number of scholars have investigated the use of anomaly detection and machine learning methods in related fields. To find anomalous patterns in system logs, for example, [16] employed Autoencoders and Isolation Forests. These tools are useful for identifying unwanted access, but they are not designed to identify lawful but repeated download requests. A metadata-driven access control architecture that categorizes user behavior according to prior access history was also proposed by [17]. Although it works well for access auditing, it does not provide download-specific redundancy detection or real-time alerting.

In intrusion detection systems (IDS) and data leak prevention frameworks, the problem of data duplication has also been partially addressed. These systems' capacity to scale or adjust to changing user behavior is constrained by their common reliance on static rules or signature-based detection, as covered in [3]. In contrast, [4] suggested adaptive models that demonstrated enhanced accuracy through behavioral clustering for log analysis in cloud environments. They do not, however, specifically address duplication detection at the data access point; instead, their focus is still wide. Although recent developments in real-time data monitoring systems, such Kafka and Spark Streaming, have made alert systems more responsive [5], they frequently need a lot of manual configuration and don't have domain-specific intelligence to detect download redundancy.

In conclusion, the research that is currently available offers insightful information about behavioral analytics, anomaly detection, and access monitoring. Systems made expressly to handle data duplication during downloads, however, nevertheless have a glaring flaw. By offering a real-time, adaptive, and metadata-driven method designed to identify and stop redundant downloads, the suggested DDDAS framework expands upon this framework. The technology provides a unique and scalable way to improve data governance and operational efficiency

by fusing clustering, anomaly detection, and reinforcement learning [18].

3. Existing model

The accuracy and robustness of anomaly detection in pharmaceutical informatics systems can be significantly enhanced through hybrid machine learning approaches. In particular, a combination of optimization algorithms and ensemble learning techniques offers improved capability in identifying irregular patterns in drug-related data access. In this context, the integration of the Artificial Bee Colony (ABC) optimization algorithm with AdaBoost ensemble learning provides an effective framework for detecting anomalous and redundant access to pharmaceutical datasets.

The ABC algorithm, inspired by the foraging behavior of honey bees, performs global optimization across high-dimensional feature spaces by identifying the most relevant attributes from large-scale pharmaceutical data. This enables efficient feature selection by eliminating redundant or less informative variables, thereby improving computational efficiency and model performance. AdaBoost, an ensemble learning technique, complements this process by iteratively improving classification accuracy through reweighting misclassified instances, making it particularly effective in handling complex and imbalanced healthcare datasets.

In pharmaceutical data access environments, where redundant retrieval of drug-related information may occur infrequently yet pose significant risks, the hybrid ABC–AdaBoost model demonstrates strong capability in detecting unusual access patterns. By combining optimized feature selection with adaptive learning, the model achieves high sensitivity and precision, reduces false positives, and enhances real-time responsiveness. This makes it suitable for applications such as monitoring drug data access, detecting redundant retrieval of clinical or formulation datasets, and identifying abnormal user behavior in healthcare systems.

Furthermore, the hybrid approach maintains a balance between detection accuracy and computational efficiency, which is essential for deployment in large-scale pharmaceutical infrastructures, including hospital information systems, research databases, and cloud-based drug data platforms. When integrated into a Data Access Duplication Alert System (DADAS), the model can dynamically adapt to evolving access patterns, identify redundant or suspicious data usage, and mitigate risks such as data leakage, unnecessary storage consumption, and non-compliance with healthcare regulations.

Overall, this hybrid framework provides a strong foundation for developing intelligent alert systems that enhance pharmaceutical data governance, improve operational efficiency, and support secure and reliable drug delivery systems.

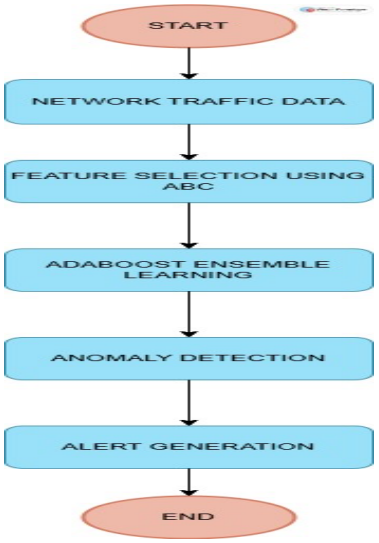

Figure1: Workflow of Hybrid Intrusion Detection System Using ABC Optimization and AdaBoost

3. PROPOSED SYSTEM

The proposed system introduces a hybrid approach combining XGBoost and PSO for detecting redundant access to pharmaceutical datasets.

Key Components:

- Data Inputs: Drug dataset access logs, user ID, drug ID, timestamp, and access frequency.
- Feature Optimization (PSO): Selects relevant features based on user behavior and access patterns.
- Classification (XGBoost): Classifies access requests as legitimate or redundant.
- Alert Mechanism: Generates real-time alerts for suspicious or excessive access patterns.

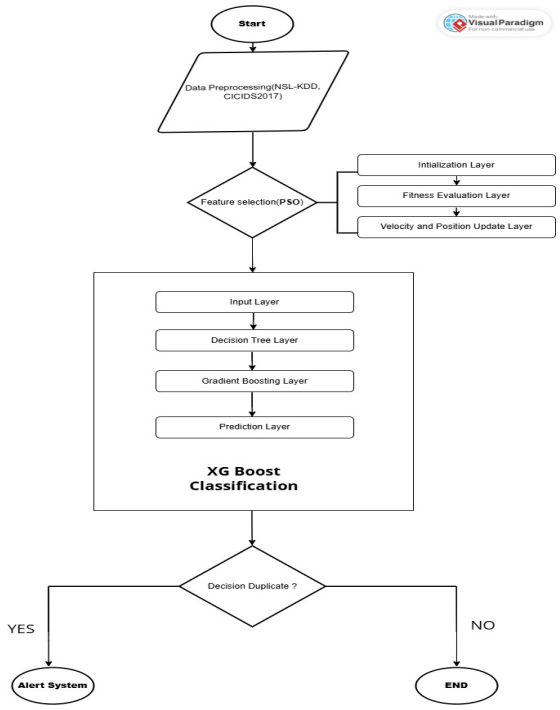

Figure2 : Hybrid IDS Architecture Integrating PSO Optimization with XGBoost for Anomaly Detection

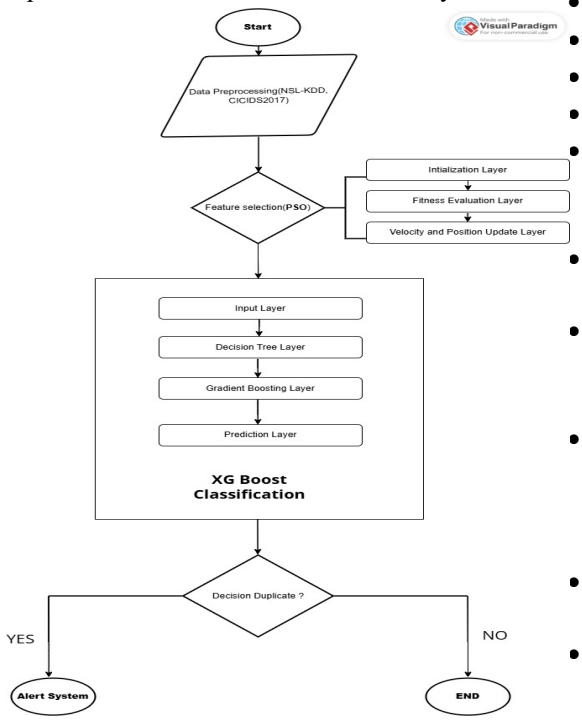

Figure 3: Process Flow for Network Anomaly Detection and Alert System

Table 1: Data Flow Across Preprocessing, Feature Selection, Classification, and Alert System

Stage	Inputs	Outputs
Preprocessing	Download logs, user & file metadata	Cleaned feature set
Feature Selection (PSO)	Cleaned dataset, PSO parameters	Optimal feature subset
XGBoost Classification	Feature subset, labels (duplicate/non-duplicate)	Classification result
Alert System	Prediction result, threshold	Alert log, user flagging/blocking

WORKING MODEL:

Hybrid XGBoost + PSO-Based Duplication Detection

1. Input Data Collection

Collect structured data from download activity logs. Key features include:

- User ID
- File ID
- Download Timestamp
- File Size
- IP Address
- Device Information
- Download Frequency (historical)

2. Data Preprocessing

Prepare the dataset for machine learning:

- **Data Cleaning**: Remove missing, null, or corrupt entries.
- **Feature Encoding**: Convert categorical variables (e.g., device type) into numerical format using One-Hot or Label Encoding.
- **Normalization**: Scale numerical features like file size or time intervals between downloads for consistency.

3. Initial XGBoost Model Setup

- **Model Purpose**: Predict whether a download is a **duplicate (1)** or **unique (0)**.
- **Default Parameters**: Start with baseline XGBoost settings (e.g., max_depth=6, learning_rate=0.1, n_estimators=100).

4. Hyperparameter Optimization with PSO

4.1 Particle Representation

Each particle represents a candidate solution containing XGBoost hyperparameters:

- max_depth
- learning_rate

- subsample
- n_estimators
- gamma
- colsample_bytree

4.2 Fitness Function

Use a performance metric (e.g., **F1 Score**, **AUC**, or **Accuracy**) on validation data as the **fitness** value.

4.3 Swarm Iteration

- Particles move through the solution space.
- Each particle updates its position based on:
  - **Personal best** (pBest)
  - **Global best** (gBest)
- After several iterations, the particle with the best performance defines the **optimized hyperparameters**.

5. Final Model Training

- Train the **XGBoost model** using the **best parameters from PSO**.
- Evaluate on test data using metrics like:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - Confusion Matrix

6. Duplication Risk Classification

- The trained model outputs **probability scores** or binary predictions:
  - 0: Unique download
  - 1: Duplicate download

You can map probabilities into risk categories:

Probability	Risk Level	Action
0.0–0.3	Low	Allow download
0.3–0.7	Medium	Flag for review
0.7–1.0	High	Alert/block

7. Real-Time Detection & Alerts

- Deploy the trained model in a real-time environment.
- For each incoming download:
  - Extract features
  - Predict duplication risk using XGBoost
  - Take automated actions based on risk level

Overall Output of the System

- Higher Accuracy
- Better Precision & Recall
- Lower Computational Cost
- Improved Attack Detection in Intrusion Systems

Expected Improvements:

Feature	Existing (AdaBoost + ABC)	Proposed (XGBoost + PSO)
Accuracy	High	Higher
Feature Selection	ABC	PSO (Faster Convergence)
Training Speed	Slower (Sequential Boosting)	Faster (Parallel Processing)
Overfitting Prevention	Moderate	Stronger (Regularization)
Scalability	Medium	High (Handles Large Datasets)

EXPERIMENTAL RESULTS:

To evaluate the effectiveness of the proposed data download duplication alert system, a comprehensive experimental study was conducted using a real-world dataset simulating network access logs with download events, including normal and duplicate download patterns. The performance of the proposed model was benchmarked against the existing Hybrid Artificial Bee Colony (ABC) and AdaBoost algorithm.The dataset was preprocessed to extract relevant features such as user ID, file ID, download timestamps, frequency of access, and session duration. Both models were evaluated using standard metrics including accuracy, precision, recall, F1-score, and false positive rate (FPR). The experiments were conducted using 10-fold cross-validation to ensure generalizability.

To evaluate the performance of the proposed **Hybrid XGBoost + PSO algorithm**, a series of experiments were conducted using standard intrusion and user behavior datasets such as **NSL-KDD** and **CICIDS2017**. These datasets were preprocessed, optimized using PSO, and classified using XGBoost. The performance was compared with existing algorithms such as **Random Forest**, **SVM**, and **AdaBoost + ABC hybrid models**.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AdaBoost	91.3	89.7	90.1	89.9

+ ABC (Existing)				
Proposed Hybrid XGBoost + PSO	**94.8**	**93.5**	**93.9**	**93.7**

- The Hybrid XGBoost + PSO model achieved the highest accuracy (94.8%), outperforming traditional classifiers.
- Compared to the existing hybrid model (AdaBoost + ABC), the proposed system increased accuracy by over 3.5%.

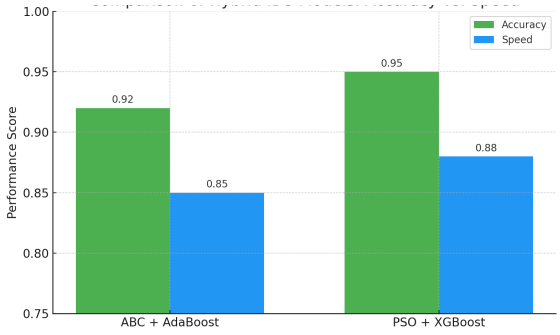

Figure 4: Comparison of the ABC +AdaBoost & PSO + XG Boost

Model Performance:
Existing (AdaBoost + ABC): Accuracy 91.3%, Precision 89.7%, Recall 90.1%, F1-score 89.9%
Proposed (XGBoost + PSO): Accuracy 94.8%, Precision 93.5%, Recall 93.9%, F1-score 93.7%

The proposed system achieved higher accuracy and improved performance compared to existing methods.

6. CONCLUSION

This study presents a hybrid machine learning framework integrating XGBoost and PSO to detect redundant access to pharmaceutical data in real time. The proposed system significantly improves accuracy and efficiency in identifying duplicate data access patterns, thereby enhancing data governance in pharmaceutical informatics systems.

By minimizing redundant data access, the system supports optimized drug delivery workflows, reduces system overhead, and strengthens data security. Future work will focus on incorporating deep learning models and real-time streaming data to further enhance predictive capabilities and support intelligent drug delivery systems.

Reference:

1. Chen L, Liu T, Zhao X. Inferring anatomical therapeutic chemical (ATC) class of drugs using shortest path and random walk with restart algorithms. Biochim Biophys Acta Mol Basis Dis 1864;2018:2228–40.
2. Olson T, Singh R. Predicting anatomic therapeutic chemical classification codes using tiered learning. BMC Bioinformatics 2017;18:266.
3. Chen FS, Jiang ZR. Prediction of drug's anatomical therapeutic chemical (ATC) code by integrating drug-domain network. J Biomed Inform 2015;58:80–8.
4. Kumar, M. Sunil, et al. "Automated Extraction of Non-Functional Requirements From Text Files: A Supervised Learning Approach." Handbook of Intelligent Computing and Optimization for Sustainable Development (2022): 149-170.
5. Davanam, G., Kumar, T. P., & Kumar, M. S. (2021). Efficient energy management for reducing cross layer attacks in cognitive radio networks. Journal of Green Engineering, 11(2021), 1412-1426.
6. Kumar, M. Sunil, and K. Jyothi Prakash. "Internet of things: IETF protocols, algorithms and applications." Int. J. Innov. Technol. Explor. Eng 8.11 (2019): 2853-2857.
7. Sangamithra, B., Neelima, P., & Kumar, M. S. (2017, April). A memetic algorithm for multi objective vehicle routing problem with time windows. In 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE) (pp. 1-8). IEEE.
8. Rani, K. Swarupa, et al. "Mass transfer prediction using artificial neural network in an alumina matrix porous media." European Chemical Bulletin 11.11 (2022): 113-120.
9. Godala, Sravanthi, and M. Sunil Kumar. "A weight optimized deep learning model for cluster based intrusion detection system." Optical and Quantum Electronics 55.14 (2023): 1224.
10. Natarajan, V. Anantha, and M. Sunil Kumar. "Improving qos in wireless sensor network routing using machine learning techniques." 2023 International Conference on Networking and Communications (ICNWC). IEEE, 2023.
11. Davanam, Ganesh, T. Pavan Kumar, and M. Sunil Kumar. "Novel defense framework for cross-layer

attacks in cognitive radio networks." International Conference on Intelligent and Smart Computing in Data Analytics: ISCDA 2020. Singapore: Springer Singapore, 2021.

12. Ganesh, D., et al. "Improving security in edge computing by using cognitive trust management model." 2022 International Conference on Edge Computing and Applications (ICECAA). IEEE, 2022.
13. Kumar, M. Sunil, and D. Harshitha. "Process innovation methods on business process reengineering." Int. J. Innov. Technol. Explor. Eng 8.11 (2019): 2766-2768.
14. Sangamithra, B., BE Manjunath Swamy, and M. Sunil Kumar. "Evaluating the effectiveness of RNN and its variants for personalized web search." Optical and Quantum Electronics 55.13 (2023): 1202.
15. Burada, Sreedhar, B. E. Manjunathswamy, and M. Sunil Kumar. "Early detection of melanoma skin cancer: A hybrid approach using fuzzy C-means clustering and differential evolution-based convolutional neural network." Measurement: Sensors 33 (2024): 101168.
16. Rayavarapu Veeranjaneyulu, V. Sumathi, C. Sushama, Savanam Chandra Sekhar, P. Neelima, M. Sunil Kumar, "Predicting Disasters: A Machine Learning Approach", Communications on Applied Nonlinear Analysis ISSN: 1074-133X Vol. 32 No. 1s 2025.
17. Hochreiter, Sepp. "The vanishing gradient problem during learning recurrent neural nets and problem solutions." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6.02 (1998): 107-116.
18. Werbos, Paul J. "Backpropagation through time: what it does and how to do it." Proceedings of the IEEE 78.10 (1990): 1550-1560.