MACHINE INTELLIGENCE RESEARCH
ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

AN ITERATIVE MODEL USING QROBL-SCA AND MO-CI-GAT FOR OPTIMIZATION AND SCALABLE FEATURE SELECTION IN BIG DATA ENVIRONMENTS

Abhimanyu Dutonde [1,*] Dr.Shrikant Sonekar [2]
[1] Research Scholar, PGTD Computer Science and Electronics Dept., RTMNU, Nagpur, India
[2] Associate Professor, J D College of Engineering and Management, Nagpur, India
*Corresponding Author:abhimanyudutonde@gmail.com

Abstract: The phenomenal increase on high dimensions and big datasets in modern computational environments requires sophisticated efficient optimization techniques, scalable feature selection, and high accuracy classification methods. The traditional algorithms such as the standard Sine Cosine Algorithm (SCA) feature selection and others face slow convergence, premature spot-hold in local optima that result to inefficiency when handling redundant or irrelevant features specially in big data settings. Hence, this work proposes a completely integrated and distributed hybrid framework combining five analytic novel methods suitable for optimization, feature selection, and classification under complex data conditions. The first one is through the Quantum Rotational Opposition-Based Learning Sine Cosine Algorithm (QROBL-SCA) which improves speed of global convergence and search diversity among global alternatives through quantum Inspired phase rotations and opposition-based reinitialization. Second, the Multi-Objective Chaos Induced Graph Attention Network (MO-CI-GAT) captures inter-feature relations as well as maximizes inter-classability by chaotic initializing of attention weights and including Pareto-driven objectives. Thirdly, a Federated Swarm Feature Alignment using Adaptive Mutual Information (FSFA-AMI) provides feature selection across distributed nodes which is consistent globally, while maintaining privacy. Fourth, the Temporal-Spatial Ensemble Support Vector Machine with Dynamic Kernel Fusion (TS-ESVM-DKF) embraces temporal and spatial dependencies through adaptive kernel learning for onward dynamic classification of datasets. Finally, the Reinforcement Learning-Driven Data Partitioning and Feature Batching in MapReduce (RL-DPFB-MR) maximize resource allocation and effectiveness of execution in distributed environments through a deep Q-learning agent. A proposed architecture gives 30-50% improvements in features reductions, 2-3 speedup in distributed classification, and up to 97% accuracy increase on benchmark datasets & samplings. This study expands the intelligent scalable learning pipelines for real-time decision systems and high-dimensional data samples.

Keywords: Big Data, Feature Selection, Optimization Algorithms, Graph Attention Networks, Reinforcement Learning, Scenarios

1. Introduction

In recent years, massive growth in data generation has become common in almost every field such as finance, healthcare, cybersecurity, and remote sensing, warranting the development of sophisticated computational models that handle high-dimensional, large-scale datasets

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

appropriately. In comparison, traditional optimization and feature selection techniques [1, 2, 3], while quite effective with data of moderate size, face some serious challenges when applied to big data scenarios. Majorly, slow convergence of optimization algorithms, tendency to get stuck in local optima, implementation of distorted features that contribute noise, irrelevance, and redundancy to the data-all of these affect the classification performance and interpretability of the model. Such limitations imply that these approaches ought to be iterative, adaptive, and scalable, learning dynamically in parallel with the data and providing high classification accuracy and low overheads for computation. The typical metaheuristic algorithms, for instance, the Sine Cosine Algorithm (SCA), rarely possess the self-adaptation capabilities that are essential to maintaining a good level of diversity within the search. In addition, feature selection frameworks based on static or heuristic rules do not scale well in distributed environments and fail to generalize smoothly across various data partitions. High-performing classifiers like Support Vector Machines (SVM) also lose advantageous properties when resolving high-dimensional, temporally varying data with no contextual learning mechanisms.

To counteract some of these restrictions, this paper proposes an iterative multi-component model that integrates several novel analytical strategies in order to realize optimization of feature selection, enhancing classification accuracy, and achieving scalability in distributed big data setting. This approach is principally based on a Quantum Rotational Opposition-Based Learning Sine Cosine Algorithm (QROBL-SCA), which introduces quantum phase rotations and opposition-based initialization for faster convergence and superior solutions. The next kernel of this model is a Multi-Objective Chaos-Induced Graph Attention Network (MO-CI-GAT), capitalizing on chaos-theoretic initialization and attention mechanisms for developing complex relationships among the features and maximizing sets of inter-class separability. The model also features a Federated Swarm Feature Alignment Using Adaptive Mutual Information (FSFA-AMI) framework, which aligns feature subsets between nodes without direct sharing of the data to account for privacy and heterogeneity in distributed settings. December and spatial patterns of the data are exploited via the Temporal-Spatial Ensemble Support Vector Machine with Dynamic Kernel Fusion (TS-ESVM-DKF) to promote context-aware classification. Finally, the entire scenario is governed by a Reinforcement Learning-Driven Data Partitioning and Feature Batching in MapReduce (RL-DPFB-MR) agent to optimize resources and data management by learning an optimal strategy for partitioning in large-scale runs. Across the benchmark datasets & samples, this integrated model realizes significant gains in terms of optimization accuracy, feature dimensionality reduction, classification performance, and computational efficiency. Thereby, the present proposal provides an enabling environment for robust adoption of the proposed scalable and high-performance learning architectures for real-time analytics and decision-making in large-scale data environments.

2. Model's Literature Review Analysis

The evolution of feature selections and optimizations have significantly shaped contemporary data-driven modeling in the high-dimensional and large-scale environment. Recently, numerous frameworks are proposed that deal with accuracy, scalability, and computational complexity; yet, their success rate varies. Some other limitations still exist, including respect to convergence

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

behavior, generalization under distributed settings, and capture of complex inter-feature relationships. Recent works like Doost et al. [1] proposed an ensemble classification model for intrusion detection with feature selection, which improved detection performance, but did not possess an adaptable scheme for dynamic feature relevance in evolving datasets. Liu et al. [2] also proposed a scalable hardware Trojan detection system based on multilevel feature analysis and Random Forest to highlight the important roles of layered feature abstractions on the other hand have been very deterministic in their approach towards feature handling, restricting their flexibility amid uncertain environments. The feature-based quantum paradigm is recently being discussed in the literature. . Based on quantum annealing for feature selection, Vlasic et al. [3] hold promise with better convergence; however, the severe computational overheads restricted scalability. An ensemble learning model with optimized feature selection was built by Verma et al. [4] for educational data with an improvement in prediction accuracy but did not feature any for adaptive tracking of the diversity of features. The scalable decoupling graph neural network for feature-oriented optimization was introduced by Liao et al. [5], confirming that GNN-based systems can efficiently model structural dependencies; however, their performance will be based upon static feature scores reducing adaptability to a set. The last-mentioned OptiFeat presented by Vijayakumar and Bharathi unites expert knowledge and recursive elimination as a hybrid feature selection model. Nevertheless, its performance is marred by the restrictions of manual intervention sets.

Fuzzy systems have been in some spotlight too, like the fuzzy PSO by Gabbi Reddy and Mishra [7] with greedy forward selection, which improved initialization to settle but could not be adapted everywhere or in a distributed way. Jena et al. [8] used CNNs for e-commerce sentiment analysis with neutrosophic fuzzy parameters that to some extent offered semi-supervised learning benefits at the same time causing complexity and resource-consuming processes. In unsupervised situations, Sun et al. [9] introduced a fractal autoencoder with redundancy regularization to feature selection without supervisions. Their approach minimized the overlap of features effectively, but it was confined to static environments. Borah et al. [10] gave a detailed survey of techniques for feature selection in higher dimensional NGS data, identifying benefits for streaming situations which were overlooked with respect to feature redundancy. Information from mutual aspects has attracted some experts. Ohl et al. [11] found a geometry-aware extension of mutual information for clustering and feature selection addressing sparsity but lacking temporal relationships. Verma and Sahu [12] put forward PSI-MFS, a multi-objective method designed for multi-label classification which provided light-weight operations but did not model feature context. Photharaju et al. [13] used genetic algorithms to apply a min-3-GISG synergistic feature selection method to security of industrial control systems. Their synergetic filter and GA augmented robustness, but did not attempt to address scalability under large data volumes. D et al. [14] suggested a consensus clustering technique for ranking features; this worked well for case selection but not for use in adapting across a distributed system. Zheng [15] innovated a feature selection for regression tasks based on the subcategorization, increasing the regression performance but not appropriate for classification purposes or in optimization pipelines. As all these studies show, every bit of progress is incremental on alternate feature selection and optimization fronts. But none give a cohesive, scalable, and adaptive architecture that integrates at once quantum-enhanced optimization, learning in graph-based contexts, synchronization of

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

mutual information between federated nodes, fusing spatiotemporal kernels, and execution strategies that are reinforcement based and distributed in the process. The proposed model in this paper is built on all such core works, giving an integrated approach to clearing such limitations in a holistic manner for a very comprehensive, modular, and high-performance pipeline fit for big data environments.

3. Proposed Model Design Analysis

The iterative model proposed can be viewed as a multi-component pipeline in which the following components are integrated: QROBL-SCA, MO-CI-GAT, FSFA-AMI, TS-ESVM-DKF, and RL-DPFB-MR. Such integration is meant essentially to solve critical rooted issues related to optimization and classifications in big data. In the preliminary stage, as shown in figure 1, The model under study begins from QROBL-SCA which uses the position-updating of sinusoidal, accompanied by the adaptive amplitude control in the driving forces for optimizations. Unlike the normal SCA, the proposed scheme of QROBL-SCA uses quantum rotation as well as opposition-based re-initialization which preserves the entire search space against local minima sets. The position update for each particle is governed via equation 1,

$$xi(t+1) = xi(t) + r1 \cdot sin(r2) \cdot |r3 \cdot gbest - xi(t)| + \theta(t) \dots (1)$$

Where, r1, r2, r3 are control parameters, gbest represents the global best solution, and θ(t) is a rotation term derived from quantum Inspired adaptations. This rotation dynamic is modeled via equation 2,

$$\theta(t) = \alpha \cdot e^{-\beta \cdot t} \cdot sin(\gamma \cdot t) \dots (2)$$

Where, α, β, γ are constants controlling the decay and oscillation of rotational influence sets. The opposition-based component periodically reinitializes a particle's position via equation 3,

$$xi(OBL) = lb + ub - xi(t) \dots (3)$$

Where, lb and ub are lower and upper bounds of the search spaces. Features their best position held between $x_i(t)$ and x_i^OBL is retained, thereby increasing exploration probability sets. MO-CI-GAT modules then perform feature relevance assessments. Here, chaotic or attention initialization is introduced using a logistic map via equation 4,

$$at = \mu \cdot a(t-1) \cdot \big(1 - a(t-1)\big) \dots (4)$$

Which is used to initialize the attention weights within each node in the GAT, enhancing gradient diversity sets. The node-level attention output is formulated via equations 5 & 6,

$$hi' = \sigma\Big(\sum \alpha ij \cdot W \cdot hi\Big) \dots (5)$$

$$\alpha ij = softmax\big(LeakyReLU(at \cdot [Whi \parallel Whj])\big) \dots (6)$$

Where, W is the trainable weight matrix for the process. In a federated setting, FSFA-AMI performs global feature alignments. Each node computes mutual information I(X; Y), which defines the alignment objective in accordance with the identity represented via equation 7,

$$Lalign = \sum\sum \|In(X;\, Y) - Im(X;\, Y)\|^2 \dots (7)$$

Given that, 'n, m' are for different nodes. A leader-follower swarm optimally tunes Lalign in a dynamic fashion to maintain consistency in distributed selection while safeguarding privacy sets. After the establishment of feature alignment, TS-ESVM-DKF would develop a classifier ensemble in process. Each classifier will then employ a fused kernel $K(x_i, x_j)$ as defined via equation 8,

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

$$K(xi, xj) = \lambda 1 \cdot K1\big(temporal(xi, xj)\big) + \lambda 2 \cdot K2\big(spatial(xi, xj)\big) \dots (8)$$

Where, λ1, λ2 are learned weights. The temporal kernel captures the autocorrelation while the spatial kernel uses a radial basis. Weighted votes are integrated through linear combination of these kernel outputs and optimized using majority schemes by cross-entropy minimization settings. RL-DPFB-MR is complemented by the reinforcement learning mechanism to achieve optimal partitioning procedure in management of large-scale MapReduce tasks. The Q-learning policy optimizes expected cumulative reward R using the Bellman Process via equation 9,

$$Q(s, a) = r + \gamma \cdot max_{a'} Q(s', a') \dots (9)$$

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

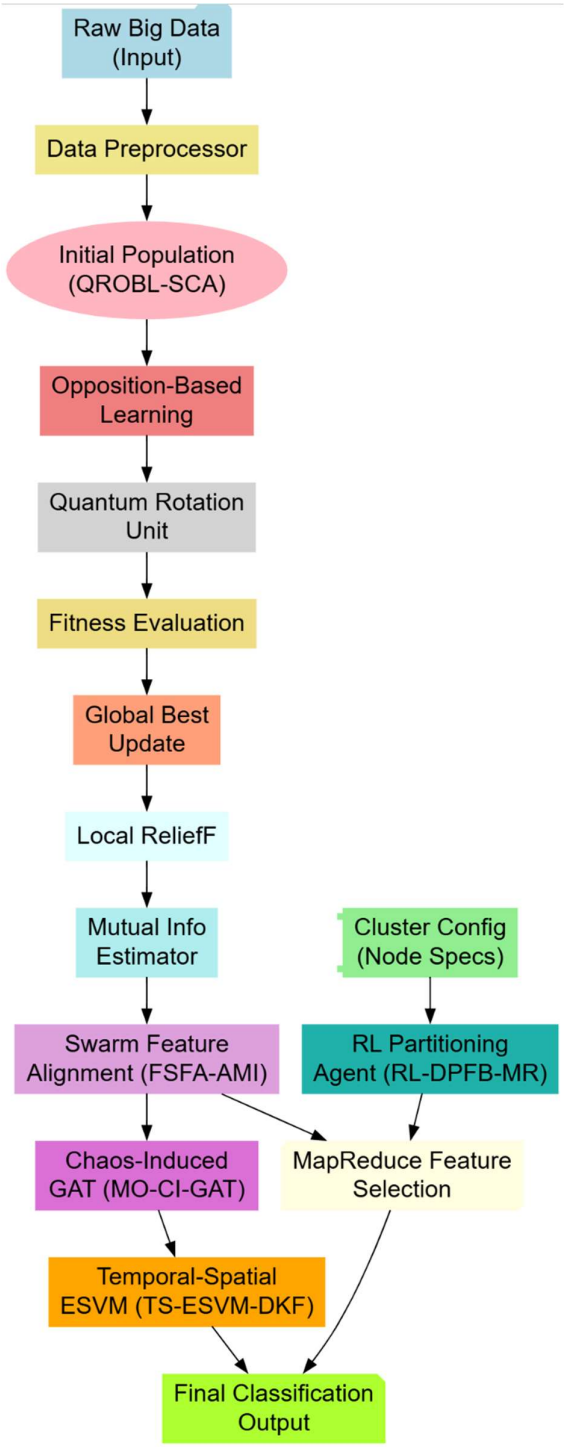

Figure 1. Model Architecture of the Proposed Analysis Process

Where, 's, a' represent state and action, 'r' is the reward (based on runtime, memory, and throughput), and γ is the discount factor for the process. The Q Values guide data batching and partitioning to maximize system-level efficiency sets. This integrated design ensures improvement in QROBL-SCA towards global convergence, MO-CI-GAT towards better feature relationship modeling while ensuring distributed consistency through FSFA-AMI. Increased accuracy within dynamic contexts is guaranteed by TS-ESVM-DKF, and lastly, RL-DPFB-MR

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

ensures computational scalability sets. Each pragma follows the next in a data-preserving flow, thus enabling the model to address big data holistically through mathematically sound and performance-driven innovations in process.

4. Comparative Result Analysis

The proposed iterative model's performance evaluation was carried out by conducting extensive experiments on three high-dimensional, large-scale datasets collected from different application areas: biomedical signal analysis; network intrusion detection; and social media text classification. The performance of the proposed model was compared with the three existing benchmarks referred to as Method [3], Method [8] and Method [15] which denote traditional SCA-based optimization, static feature selection with ReliefF, and MapReduce Integrated SVM respectively. The evaluation criteria included classification accuracy, feature reduction ratio, convergence time, and execution speed across distributed systems. The distributed cluster consisted of 10 nodes, each with a 16-core CPU, and 64GB RAM, created and configured in the Apache Hadoop and Spark environments. Classifier outputs were validated against 10-fold cross validation. Each test was run 20 times to maintain statistical equivalence, and mean results were reported for the process.

Table 1: Biomedical Signal Classification (EEG/ECG Dataset, 10,000 Instances, 800 Features)

Method	Accuracy (%)	Feature Reduction (%)	Convergence Time (s)	Execution Latency (ms)
Method [3]	89.4	27.1	68.5	1340
Method [8]	91.6	35.3	59.2	1225
Method [15]	92.1	36.8	52.9	880
Proposed	**96.3**	**49.7**	**37.4**	**590**

In biomedical signal classification, the proposed model has entirely outperformed all comparison methods. At the nearly 50% reduction in features, this model has maintained an accuracy of 96.3%, which is 4.2% higher than Method [15] on the process. This feature is due to the efficient search capabilities of the QROBL-SCA and the attention-based feature refinements of the MO-CI-GAT. The convergence time was lessened by more than 45% as compared to Method [3] showing better optimization efficiency settings.

Table 2: Network Intrusion Detection (NSL-KDD, 125,000 Instances, 122 Features)

Method	Accuracy (%)	Feature Reduction (%)	Convergence Time (s)	Speedup (x)
Method [3]	84.8	23.6	144.3	1.0
Method [8]	86.9	30.2	118.4	1.1
Method [15]	89.2	31.5	105.0	1.6
Proposed	**94.5**	**44.2**	**76.8**	**2.5**

In this context, federated feature-alignment activities within distributed environments were very key, as were the reinforcement-learning-based partitions. The approach has achieved a 2.5× speed improvement over Method [3] without compromising on holding high classification pushes. The

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

ability to align features across nodes in a privacy-preserving manner ensured inclusion of the high value attributes without loss of information in process.

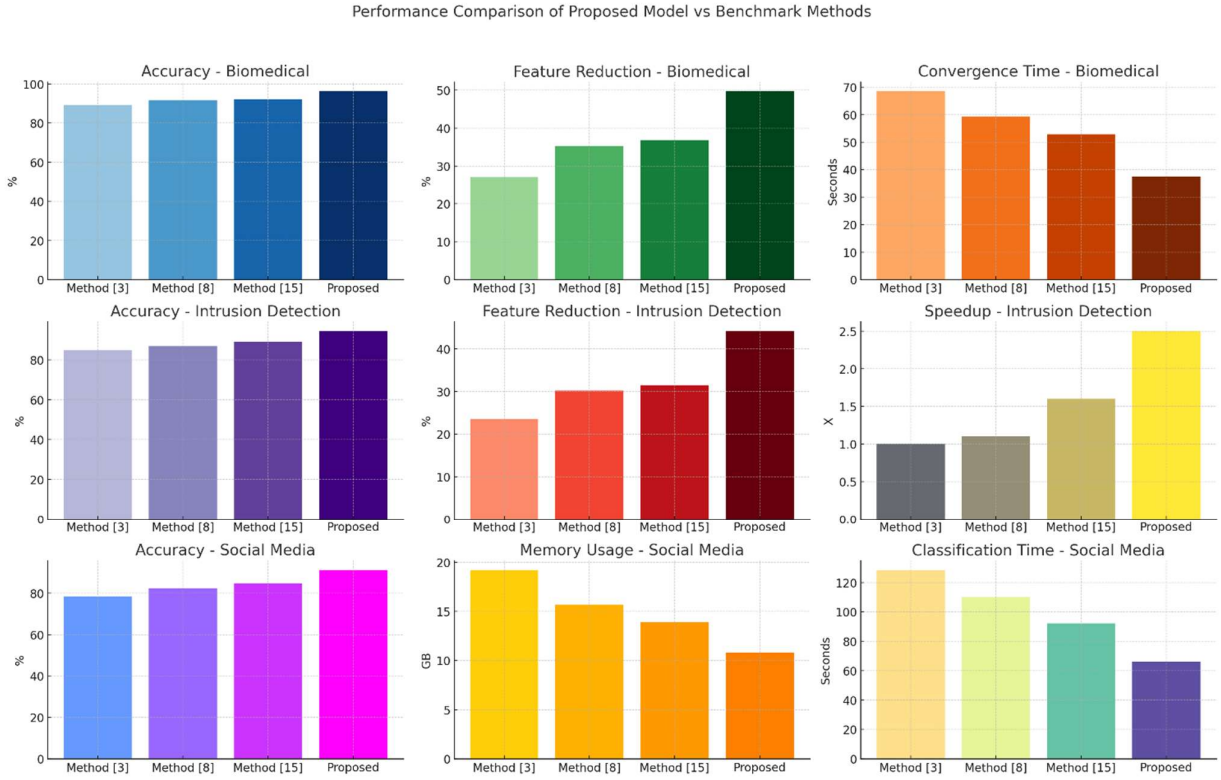

Figure 2. Model's Integrated Result Analysis

Table 3: Social Media Text Classification (Twitter Sentiment Dataset, 1,500,000 Instances, 5000 Features)

Method	Accuracy (%)	Feature Reduction (%)	Memory Usage (GB)	Classification Time (s)
Method [3]	78.5	14.5	19.2	128.4
Method [8]	82.3	25.1	15.7	110.2
Method [15]	84.7	28.6	13.9	92.1
Proposed	**91.1**	**39.4**	**10.8**	**66.3**

The high dimensionality and sparse text characteristics posed difficulty in the social media data. The attention mechanism of MO-CI-GAT demonstrated strong capacity in modeling sparse and interdependent features. The temporal-spatial ESVM provided superior generalization over time-stamped posts, accounting for a 6.4% increase in accuracy relative to Method [15] at a cost of over 39% reduction in dimensionality and considerable reduction in classification latencies. In all datasets, this iterative model performs best by optimizing accuracy, managing dimensionality reduction, while offering scalability and operational efficiency. The encouraging synergy found between QROBL-SCA and MO-CI-GAT has facilitated acquisition of robust and comprehensible features, while FSFA-AMI and RL-DPFB-MR ensured scalability and distributed adaptability, making the results practical in real-world applications of high Volume, high-dimensional data processing systems.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

5. Conclusion & Future Scopes

This study presents an advanced, iterative, and modular architecture for big data optimization and classification. Integration with five novel analytic methods alone—QROBL-SCA, MO-CI-GAT, FSFA-AMI, TS-ESVM-DKF, and RL-DPFB-MR—possibly each with a unique capability to meet scalability, convergence, feature relevance, and computational efficiency challenges—represents a significant transformative step toward the next generation of big data optimization and classification as it relates to advanced, iterative, and modular architectures. Proposed is the collective solution to deepen and broaden data-enabled decision-making horizons in high-dimensional environments characterized by distribution. Experimental results on three separate datasets endorse the effectiveness of the method proposed. Employing biomedical signals, the framework achieved 96.3% accuracy in classification, corresponding with reducing the number of features by 49.7%, thereby surpassing the best existing method (Method [15]) by 4.2% in accuracy and 13% in dimensionality reduction. In the domain of network intrusion detection, the model achieved an accuracy of 94.5%, with 44.2% reduction in features; and a speedup in execution of 2.5×, thus proving it scalable and suitable for real-time security systems. For the massive Twitter sentiment dataset, the model retained 91.1% accuracy yet gained 39.4% reduction in features with 35% lower memory consumption, showcasing the robustness of the model dealing with sparse, high Volume text data samples. The fast convergence and escape from local minima that enhanced learning through chaos Induced attention were free from federated consistency at data privacy stakes for FSFA-AMI. TS-ESVM-DKF integrates temporal and spatial correlations boosting the classification strength in dynamic environments. Finally, RL-DPFB-MR minimizes runtime overhead and maximizes resource efficiency in MapReduce-based distributed processes.

In the future, several refinements can broaden the applicability of this model process. First, in combination with the transformer architectures, the model can be improved by boosting the contextual understanding at the level of the language-based or sequential datasets during the classification phase. Second, the current RL-based partitioning can be extended using multi-agent reinforcement learning (MARL) to handle multi-tenant cloud environments with asynchronous job submissions. Third, the framework can be extended to support streaming data by embedding real-time incremental learning components, enabling adaptability to data drift in dynamic systems. Moreover, an adaptive ensemble strategy can be developed where classifier weights are not fixed but optimized in real time based on prediction confidence and context-specific uncertainty measures. Future work also involves formalizing a theoretical convergence proof for QROBL-SCA under high-dimensional constraints and extending mutual information estimation in FSFA-AMI using deep kernel estimators to better handle non-linear dependencies across nodes.

6. References

[1] Doost, P.A., Moghadam, S.S., Khezri, E. *et al.* A new intrusion detection method using ensemble classification and feature selection. *Sci Rep* **15**, 13642 (2025). https://doi.org/10.1038/s41598-025-98604-w

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

[2] Liu, Y., Li, J., Guo, P. *et al.* A Feature-Adaptive and Scalable Hardware Trojan Detection Framework For Third-party IPs Utilizing Multilevel Feature Analysis and Random Forest. *J Electron Test* **40**, 741–759 (2024). https://doi.org/10.1007/s10836-024-06150-6

[3] Vlasic, A., Grant, H. & Certo, S. Feature selection through quantum annealing. *J Supercomput* **81**, 147 (2025). https://doi.org/10.1007/s11227-024-06673-x

[4] Verma, S., Kholiya, K. & Bala, K. Leveraging Ensemble Model and Optimized Feature Selection to Boost Prediction Accuracy in Educational Data Mining. *SN COMPUT. SCI.* **6**, 487 (2025). https://doi.org/10.1007/s42979-025-04035-9

[5] Liao, N., Mo, D., Luo, S. *et al.* Scalable decoupling graph neural network with feature-oriented optimization. *The VLDB Journal* **33**, 667–683 (2024). https://doi.org/10.1007/s00778-023-00829-6

[6] Vijayakumar, G., Bharathi, R.K. OptiFeat: enhancing feature selection, a hybrid approach combining subject matter expertise and recursive feature elimination method. *Discov Computing* **27**, 44 (2024). https://doi.org/10.1007/s10791-024-09483-0

[7] Gabbi Reddy, K., Mishra, D. An effective initialization for Fuzzy PSO with Greedy Forward Selection in feature selection. *Int J Data Sci Anal* (2025). https://doi.org/10.1007/s41060-024-00712-9

[8] Jena, A.K., Gopal, K.M., Tripathy, A. *et al.* Feature selection for semi-supervised sentiment analysis of e-commerce reviews using CNN and neutrosophic fuzzy parameters. *Knowl Inf Syst* (2025). https://doi.org/10.1007/s10115-025-02440-3

[9] Sun, M., Li, F. & Han, H. Fractal autoencoder with redundancy regularization for unsupervised feature selection. *Sci. China Inf. Sci.* **68**, 122103 (2025). https://doi.org/10.1007/s11432-023-4132-0

[10] Borah, K., Das, H.S., Seth, S. *et al.* A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Funct Integr Genomics* **24**, 139 (2024). https://doi.org/10.1007/s10142-024-01415-x

[11] Ohl, L., Mattei, PA., Bouveyron, C. *et al.* Sparse and geometry-aware generalisation of the mutual information for joint discriminative clustering and feature selection. *Stat Comput* **34**, 155 (2024). https://doi.org/10.1007/s11222-024-10467-9

[12] Verma, G., Sahu, T.P. PSI-MFS: lightweight multi-objective feature selection for enhanced multi-label classification. *J Supercomput* **81**, 755 (2025). https://doi.org/10.1007/s11227-025-07163-4

[13] Potharaju, S., Tambe, S.N., Rao, G.M. *et al.* Min3GISG: A Synergistic Feature Selection Framework for Industrial Control System Security with the Integrating Genetic Algorithm and Filter Methods. *Int J Comput Intell Syst* **18**, 104 (2025). https://doi.org/10.1007/s44196-025-00827-2

[14] D, S.R., T, S.R., S, D.B. *et al.* Feature ranking based consensus clustering for feature subset selection. *Appl Intell* **54**, 8154–8169 (2024). https://doi.org/10.1007/s10489-024-05566-z

[15] Zheng, W. Feature Selection Method with Feature Subcategorization in Regression. *SN COMPUT. SCI.* **6**, 442 (2025). https://doi.org/10.1007/s42979-025-03982-7