

AI-Driven Big Data Classification: A Taxonomy of Learning Paradigms, Scalability Approaches, and Emerging Research Directions

¹Arun K Revanth, ²Dr. K. Pazhanikumar

¹Research Scholar, Department of Computer Science, S. T. Hindu College, Nagercoil,
Affiliated to Manonmaniam Sundaranar University, Tirunelveli, India.

¹E-Mail: revanth@sthinducollege.com

²Head & Assistant Professor, Department of Computer Science, S. T. Hindu College, Nagercoil,
Affiliated to Manonmaniam Sundaranar University, Tirunelveli, India.

²E-Mail: skpk73@gmail.com

Abstract

The explosion of data from connected devices, digital platforms and sensor networks has highlighted the limitations of traditional classification algorithms and has driven the need for artificial intelligence (AI) methods that can scale to handle the volume, velocity and variety of data. This survey presents a comprehensive review of the current state of AI-enabled big data classification along three complementary axes: learning paradigms, scalability strategies and emerging research directions. We propose a taxonomy of existing methods into supervised and instance-based learning, deep and hybrid architectures, and imbalance-aware learning based on resampling. The infrastructural dimension of the problem is then discussed in relation to distributed computing frameworks such as MapReduce and Apache Spark MLlib, dimensionality-reduction pipelines and grid-based indexing schemes that keep classification tractable at scale. The surveyed application areas are healthcare and Internet of Things (IoT) systems, network intrusion detection, genomics, environmental monitoring, supply chains and multimodal emotion recognition. A comparison table compares representative techniques on method, domain and trade-offs, and a research-gap table identifies open issues in interpretability, cross-domain transfer, and energy-aware computation. The paper ends with a discussion of promising directions, including explainable hybrid pipelines, federated and streaming classification, and standardized benchmarking for large-scale artificial intelligence systems.

Keywords: Big Data Classification; Artificial Intelligence; Class Imbalance; Distributed Computing; Feature Selection; Apache Spark; Deep Learning; Scalability.

Introduction

Data volumes across almost all industries healthcare, finance, telecommunications, manufacturing, social media are growing at a rate that outpaces the capacity of traditional statistical and machine-learning tools that are meant for small, well-curated data sets. Big data is frequently discussed in terms of volume, velocity, variety, and increasingly veracity and value, each imposing different demands on classification systems that must label instances quickly, accurately, and at scale [3]. Classical classifiers like decision tree models or shallow neural networks were built under the implicit assumption of centralized, memory-resident data. Big data classification demands

architectures that can be distributed across clusters, updated incrementally as new data arrives, and tolerant to noise and missing values.

The standard toolkit for filling this gap is artificial intelligence, and deep learning in particular. In order for a single pipeline to handle terabytes of data without failing due to memory or time constraints, AI-enabled classification pipelines are rapidly integrating feature engineering, ensemble learning, and neural architectures with distributed processing engines [4]. However, the research on this topic is still dispersed: some articles concentrate only on algorithmic innovation, like novel deep architectures or hybrid feature-selection schemes, while others concentrate on the systems side, such as developing grid-indexed storage or optimizing Spark-based pipelines. Although the real-world success of a categorization system depends on both, few surveys make an effort to link these two strands.

This paper attempts to bridge that gap by proposing a taxonomy that treats AI-enabled big data classification along two orthogonal axes. The first axis relates to the learning paradigm: whether a system uses deep and hybrid architectures, classical supervised learning, or specific methods for managing imbalanced classes one of the most enduring challenges in big real-world datasets [11], [12]. The distributed computing frameworks, feature-selection techniques, and indexing systems that make classification computationally possible after data no longer fits comfortably on a single computer make up the second axis, which is scalability [1], [2], [5].

The remainder of this survey is organized as follows. Section 2 discusses the background and characteristics that distinguish big data classification from conventional classification. Section 3 develops a taxonomy of learning paradigms. Section 4 reviews scalability strategies. Section 5 surveys application domains ranging from healthcare to genomics to environmental monitoring. Section 6 offers a comparative table of representative techniques, and Section 7 presents a research-gap table. Section 8 discusses emerging directions, and Section 9 concludes the paper.

2. Background: Why Big Data Classification Is Different

Big data classification is different from traditional classification in a number of significant ways. First, dimensionality is frequently severe. For example, genomic datasets may include thousands of features but relatively few samples, a situation that is difficult for traditional algorithms to manage without rigorous feature selection [7]. Second, naive classifiers are biased toward the majority class because class distributions are often skewed, with the class of interest fraudulent transactions, rare diseases, and network intrusions representing a small portion of the total dataset [11], [12]. Third, algorithms with quadratic or worse time complexity become unfeasible due to the sheer volume of data, necessitating a shift toward distributed or approximate solutions [1].

The tools and methods utilized across the big data analytics pipeline, from intake and storage to modeling and deployment, have been cataloged in recent thorough studies [3]. Since model accuracy in a research setting often does not translate directly into production performance once latency, drift, and infrastructure constraints are introduced, complementary work has specifically focused on the practical challenges of deploying machine-learning models end-to-end in big data environments [4]. These more general surveys offer helpful background, but they typically consider

categorization as one of many tasks in the big data pipeline rather than as a topic worthy of a specific taxonomy a gap this study seeks to bridge.

The fact that "big data classification" is a series of linked issues that vary by domain rather than a single issue adds another layer of complexity. A genomics lab categorizing tumor subtypes from a limited number of high-dimensional microarray samples has a radically different data profile high-velocity, streaming, adversarial than a network security team classifying intrusion attempts. As a result, any taxonomy of the field must take into account both the shared scalability architecture that underpins the majority of contemporary pipelines and the algorithmic variety utilized to handle these various profiles. This is precisely the organizing approach used in Sections 3 and 4.

3. Taxonomy of AI Learning Paradigms in Big Data Classification

3.1 Supervised and Instance-Based Learning

Supervised classification remains the main approach in most big data pipelines, and instance-based methods such as k-nearest neighbors (KNN) are still revisited and improved for large-scale settings. Recent work has proposed enhanced versions of KNN, which alter the distance weighting and neighbor selection to improve performance on large, noisy data sets, while keeping the computational cost reasonable. This shows that even algorithms that are decades old remain relevant when they are carefully adapted to modern data properties [13]. These base classifiers are often used as building blocks of ensemble methods such as bagging, boosting and stacking, to reduce variance and increase robustness to noisy or inconsistent labels, a problem that becomes more pronounced as the dataset size increases and manual verification of each instance becomes impractical [19]. Instance-based and ensemble approaches are attractive first-line tools before more complex architectures are considered due to their relative interpretability and relatively low training cost, especially in domains where computational budgets or explainability requirements are strict, such as clinical decision support.

3.2 Deep and Hybrid Architectures

Deep learning architectures provide an alternative to supervised and instance-based methods for the learning of hierarchical feature representations, which can be particularly useful when the feature space is extremely high dimensional or highly non-linear. Incremental deep learning techniques, such as deep multiple-layer perceptrons trained incrementally on large datasets, have shown to improve classification accuracy on benchmark big data collections and avoid the memory bottlenecks of loading the whole dataset in one pass of training [8]. Another important thread in this paradigm is hybrid architectures that combine AI-driven feature selection with downstream machine-learning classifiers; by filtering the feature space before classification begins, these hybrid pipelines reduce both computational cost and the risk of overfitting on high-dimensional data, an approach shown to improve accuracy across several classification benchmarks [6]. Ensemble deep architectures take it a step further by combining dynamic classifier selection and oversampling techniques to develop systems that adapt their features, but also their model composition to the nature of the incoming data batch [9]. Overall, these approaches suggest that the best performing big data classification systems tend to combine multiple learning strategies rather than rely on any

single algorithm, a trend likely to continue as hardware accelerators make deeper and more complex ensembles computationally affordable at scale.

3.3 Resampling and Imbalance-Aware Learning

Class imbalance is perhaps the single most cited obstacle in the reviewed literature, appearing across domains as diverse as network security, genomics and industrial IoT. The basic problem is this: when one class is much more numerous than another, classifiers trained on the raw data will tend to predict the majority class, giving a high overall accuracy but completely failing on the minority class, which is often the class we really care about. This is handled by a large body of recent work, which does so through resampling. Synthetic oversampling techniques, most notably variants of the Synthetic Minority Over-sampling Technique (SMOTE), create artificial minority-class instances to rebalance the training distribution. Hybrid variants, combining oversampling with ensemble dynamic classifiers, have reported strong gains in minority-class recall on large-scale benchmark datasets [9]. Intrusion-detection systems, where harmful traffic is uncommon compared to benign traffic but bears a disproportionately large cost when misclassified, have been strengthened by the employment of similar SMOTE-based balancing in conjunction with iterative partitioning filters [18]. When working with the small, high-dimensional sample sizes typical of genomic studies, kernel density estimation has been suggested as an alternative basis for synthetic sampling in genomics, producing minority-class samples that better respect the underlying data distribution than naive interpolation-based approaches [20]. A more comprehensive analysis of resampling techniques has arranged these methods from a data-centric viewpoint, differentiating between over-sampling, under-sampling, and hybrid strategies and pointing out that no single resampling strategy is dominant across all domains; instead, the best option is largely dependent on dataset size, feature dimensionality, and the cost asymmetry between misclassification types [12]. Skewed feature distributions can impair classifier performance even in cases where class balance is otherwise reasonable, according to complementary empirical research that focuses on the effects of imbalanced features rather than imbalanced classes [11]. This suggests that imbalance-aware learning must consider more than just class labels.

4. Scalability Strategies for Big Data Classification

4.1 Distributed Computing Frameworks

One of the first and most significant methods for making machine-learning tasks manageable on commodity clusters was the MapReduce programming model. Its use in feature selection for big data classification showed that evolutionary search methods, which are typically thought to be too computationally costly for large feature spaces, could be efficiently parallelized across distributed nodes [1]. This early work established a pattern that subsequent research has continued to build upon: many approaches break down existing algorithms into map and reduce phases that can run independently throughout a cluster, as opposed to creating algorithms from scratch for distributed settings. Because of their in-memory processing paradigm, which eliminates the frequent disk I/O that made MapReduce rather inefficient for iterative algorithms, Apache Spark and its machine-learning library, MLlib, have since become a commonly used replacement for MapReduce for iterative machine-learning workloads. The systems-level configuration of a distributed pipeline is

just as important as the choice of learning algorithm itself, as demonstrated by recent work optimizing Spark MLlib pipelines for large-scale predictive modeling. Careful tuning of partitioning, caching, and hyperparameter search strategies can yield significant improvements in both training time and predictive accuracy compared to default configurations [5].

4.2 Feature Selection and Dimensionality Reduction

In large data classification, feature selection has two functions: it lowers computational cost by reducing the input space, and it frequently increases accuracy by eliminating redundant or noisy qualities that would otherwise weaken a classifier's learning signal. When applied to high-dimensional classification tasks, hybrid AI-driven feature-selection frameworks that integrate multiple selection criteria filter, wrapper, and embedded methods have been demonstrated to outperform any single selection strategy in isolation, especially when paired with downstream ensemble classifiers [6]. A recent thorough analysis of the microarray feature-selection landscape lists dozens of selection strategies and their relative strengths, noting that hybrid and ensemble selection methods are increasingly dominating the field over single-criterion filters [7]. This is especially true in fields like genomics, where microarray datasets frequently contain far more features than samples. This emphasis is echoed by broader assessments of real-world large data machine-learning deployment, which note that dimensionality reduction often makes the difference between a pipeline that can be retrained within an operating time budget and one that cannot [4].

4.3 Indexing and Grid-Based Approaches

Scalability can be handled at the data-organization level in addition to algorithmic and feature-level solutions. For low-occupancy or sparsely populated big data collections, where the majority of the space contains no data points at all, grid-based indexing techniques divide the feature space into a structured grid, enabling classification queries to be answered by looking only at the pertinent grid cells rather than the entire dataset [2]. The algorithmic techniques covered in Sections 3 and 4.1–4.2 are complemented, not replaced, by such indexing schemes. An effective grid index can significantly reduce the amount of data that a distributed classification algorithm must process initially, thereby pre-filtering the search space before more computationally costly learning routines are applied.

5. Domain-Specific Applications

5.1 Healthcare and Internet of Things (IoT)

Big data categorization in healthcare is increasingly using streams produced by IoT sensors and connected medical equipment instead of static, retrospectively collected statistics. Combining feature optimization with classification models can increase diagnostic precision while maintaining responsiveness for near-real-time monitoring applications, as demonstrated by a hybrid optimized framework created for IoT-based healthcare systems [10]. Because misclassifications have direct clinical implications, precision is crucial in this field. This has led to parallel research on explainable AI for identifying discrepancies in cancer classification datasets, with the goal of identifying instances in which a model's confident prediction may actually reflect noisy or contradictory labeling rather than a true pattern in the underlying biology [19].

5.2 Cybersecurity and Intrusion Detection

Another area where big data categorization must function in highly unbalanced and aggressive environments is network security. For network information security control, trigonometric-function-based modeling techniques have been put forth, providing a different mathematical foundation for depicting traffic patterns that can supplement traditional statistical features [15]. In order to address the extreme scarcity of malicious traffic samples compared to benign traffic, intrusion-detection systems designed specifically for big data environments have combined gravitational search optimization with SMOTE-based balancing filters, reporting high-accuracy detection even for rare attack categories that conventional classifiers frequently miss [18].

5.3 Genomics and Bioinformatics

The exceptionally high complexity of genomic classification problems in relation to sample size makes them an ideal testing ground for resampling and feature-selection studies. Imbalance-aware sampling work that employs kernel density estimation to provide more distribution-faithful synthetic minority samples for genomic classification tasks [20] directly interacts with the microarray feature-selection literature covered in Section 4.2 [7]. Given the high stakes of clinical decision-making based on genomic evidence, explainable AI algorithms applied to cancer data categorization further demonstrate how genomics researchers are increasingly prioritizing model openness alongside raw accuracy [19].

5.4 Environmental, Energy, and Supply Chain Applications

Big data categorization approaches have been used in fields related to physical and logistical systems outside of healthcare and security. In order to assist more focused urban planning actions, urban heat island monitoring research has used big data analysis tools to categorize and forecast temperature anomalies across metropolitan zones [16]. Fuzzy logic-based methods based on big data techniques have been applied in supply chain management to evaluate circularity in food supply chains, categorizing supply chain configurations depending on how well they adhere to circular economy principles [14]. These uses demonstrate how the basic classification methods covered in Sections 3 and 4 extend far beyond the security and healthcare fields that are most frequently covered in the AI literature.

5.5 Multimodal and Emotion Recognition Systems

Finally, multimodal emotion identification is an area where big data classification must incorporate several data kinds, such as text, audio, and visual signals, into a single pipeline. Multimodal emotion recognition specifically for big data contexts has been advanced by recent work combining prompt engineering with deep adaptive learning, demonstrating that large-scale pretrained models can be adapted to classification tasks that require fusing signals across modalities without requiring separate models for each input type [17]. The foundation-model paradigm, which has come to dominate other fields of AI research, and big data classification are beginning to converge more broadly, according to this line of work.

6. Comparative Analysis

Table 1 highlights twelve example strategies from the works reviewed above, contrasting the underlying method, application domain, primary strength, and a major limitation. This comparison

reveals a number of trends. First, while distributed computing-based methods like Spark MLlib optimization and MapReduce feature selection offer high scalability, they frequently require significant engineering work to properly configure for a given workload [1], [5]. Second, minority-class performance is routinely improved by resampling-based methods; nevertheless, when oversampling ratios are aggressive, these methods may generate artificial artifacts that diminish interpretability [9], [20]. Third, compared to more straightforward baselines like tuned KNN, hybrid and ensemble approaches typically yield the best reported accuracy figures across domains, but at the expense of longer training times and less transparency [6], [13].

TABLE I

Comparison of Representative AI-Enabled Big Data Classification Techniques

Ref.	Technique	Domain	Key Strength	Limitation
[1]	MapReduce evolutionary feature selection	General big data	Parallelizes expensive search across clusters	Disk-based I/O limits iterative training speed
[2]	Grid-index classification	Low-occupancy big data	Reduces search space via spatial indexing	Less effective on dense, high-dimensional data
[5]	Spark MLlib optimization	General predictive modeling	In-memory processing speeds iterative training	Performance sensitive to cluster tuning
[6]	Hybrid AI-driven feature selection	High-dimensional classification	Combines multiple criteria for higher accuracy	Added pipeline complexity and tuning overhead
[7]	Microarray feature-selection survey	Genomics	Broad comparative view of selection strategies	No single method identified as universally best
[8]	Incremental deep MLP	General big data	Avoids memory bottlenecks via incremental training	Deep models remain relatively opaque
[9]	HSMOTE + Ensemble Dynamic Classifier	Imbalanced big data	Strong minority-class recall gains	Synthetic samples may introduce noise
[10]	Hybrid optimized framework	IoT healthcare	Balances precision with near-real-time response	Framework complexity may hinder deployment
[13]	Enhanced KNN models	General classification	Simple, interpretable, low training cost	Scales poorly on very large datasets
[14]	Fuzzy circularity model	Food supply chain	Domain-specific interpretability	Limited generalizability beyond supply chains

Ref.	Technique	Domain	Key Strength	Limitation
[18]	Gravitational search + SMOTE-IPF	Network intrusion detection	High accuracy on rare attack classes	Computationally intensive optimization step
[20]	KDE-based synthetic sampling	Genomic imbalance	Distribution-faithful synthetic samples	Higher computational cost than plain SMOTE

7. Research Gaps

There are still a number of gaps in the reviewed literature despite significant advancements. These gaps are arranged per research area in Table 2, along with the related possibility for future work and the present limitation. The conflict between interpretability and accuracy is a recurrent theme: the best-performing deep and hybrid architectures are often the least transparent, which is problematic in regulated industries like healthcare and finance where non-technical stakeholders need to be able to understand model decisions [19]. Cross-domain generalization is a second recurrent theme. The majority of the reviewed techniques were validated on a limited number of benchmark datasets within a single domain, raising the question of whether, for instance, a resampling strategy validated on genomic data would transfer effectively to network security data with very different feature distributions[11],[12].A third gap relates to the energy and computational cost of scalability strategies themselves. Despite the growing popularity of cluster-scale training, surprisingly little of the reviewed literature discusses the energy consumption or computational footprint of the distributed pipelines it suggests[4],[5].

TABLE

II

Identified Research Gaps in AI-Enabled Big Data Classification

Research Area	Current Limitation	Research Opportunity
Interpretability of hybrid/deep models	High accuracy often comes with reduced transparency, limiting adoption in regulated domains	Inherently interpretable hybrid architectures or tailored post-hoc explanation methods
Cross-domain generalization	Most techniques are validated within a single domain or dataset	Systematic cross-domain benchmarking of feature-selection and resampling strategies
Energy-aware scalability	Distributed pipelines rarely report energy or computational cost	Metrics and benchmarks for energy-efficient big data classification
Streaming and concept drift	Many resampling and feature-selection methods assume static, batch data	Adaptive methods that handle evolving class distributions in real time

Research Area	Current Limitation	Research Opportunity
Standardized evaluation	Comparisons across papers use inconsistent metrics and datasets	Shared benchmark suites and reporting standards for the field
Small-sample, high-dimensional settings	Genomic and microarray classification still struggles with very small sample sizes	Improved synthetic sampling and transfer learning tailored to small-N, high-dimensional data

8. Emerging Research Directions

Building on the gaps mentioned above, a number of avenues seem particularly intriguing for further research. As hybrid architectures that incorporate feature selection, resampling, and deep classification become more sophisticated and challenging to audit, explainability is expected to shift from a secondary consideration to a design requirement. The gap between the accuracy increases reported in the literature and the transparency required by regulators and practitioners in healthcare and finance may be filled by methods that reveal which attributes or synthetic samples most affected a particular prediction.

A second potential approach is feature selection and adaptive, streaming-aware resampling. A large portion of the imbalance-aware literature examined in this survey makes the assumption that the entire dataset is accessible beforehand, enabling the computation of resampling ratios and particular features once prior to training. However, real-world large data systems often run on continuous streams where class distributions change over time a process called concept drift. Resampling and feature-selection approaches would be even more beneficial for production deployment in areas like network security and IoT monitoring if they were extended to function incrementally, adjusting their parameters as new data comes.

A third approach incorporates energy-aware and edge-deployable categorization. The energy cost of training and retraining classification models increases in tandem with the further scaling of distributed frameworks like Spark to larger clusters. Big data classification systems could be made more sustainable by research that simultaneously optimizes for classification accuracy and computational or energy cost rather than treating cost as an engineering afterthought. This is especially important as these systems are increasingly deployed on resource-constrained edge devices in Internet of Things contexts.

Finally, uniform benchmarking remains an unexplored but critical direction. It is challenging to compare methodologies directly because the publications evaluated in this survey involve datasets from many domains, such as supply chains, genomics, network traffic, and urban sensor networks, and claims of superiority are frequently domain-specific rather than generic. Instead of depending on the more disjointed, single-domain assessments that currently predominate the literature, the field would be able to more thoroughly evaluate which learning paradigms and scalability strategies truly generalize with a shared benchmark suite covering multiple domains and imbalance profiles.

9. Conclusion

This survey structured the rapidly developing literature on AI-enabled large data classification into two complementary axes: learning paradigms and scalability methodologies. We followed a path down the paradigm axis from traditional supervised and instance-based approaches to deep and hybrid architectures to specialized resampling strategies intended to rectify the class imbalance present in almost all of the application domains we looked at. We examined distributed computing frameworks, feature-selection pipelines, and indexing strategies that collectively make classification computationally possible if datasets surpass a single machine's capacity inside the scalability axis. Even while each domain presents its own data features and constraints, application fields spanning from healthcare and cybersecurity to genomics, environmental monitoring, and multimodal emotion recognition show that these techniques generalize well beyond any one discipline. Despite significant advancements, interpretability, cross-domain generalization, energy efficiency, and standardized evaluation are still unresolved issues, as shown by the comparative and research-gap tables in Sections 6 and 7. A logical and essential next step for this field of study is to fill these gaps with explainable hybrid architectures, streaming-aware resampling, energy-conscious distributed training, and shared benchmarks.

Reference

- [1] Peralta, Daniel, et al. "Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach." *Mathematical Problems in Engineering*, vol. 2015, no. 1, 2015.
- [2] Zhao, Haiyan, and Shuangxi Li. "Intelligent Classification Method of Low Occupancy Big Data Based on Grid Index." *Scientific Programming*, vol. 2021, no. 1, 2021.
- [3] Shahnawaz, M., and M. Kumar. "A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques." *ACM Computing Surveys*, vol. 57, no. 8, 2025, art. 196. DOI:10.1145/3718364.
- [4] El-Sayed, A., M. Abougabal, and S. Lazem. "Practical Big Data Techniques for End-to-End Machine Learning Deployment: A Comprehensive Review." *Discover Data*, vol. 3, 2025, art. 11. DOI:10.1007/s44248-025-00029-3.
- [5] Theodorakopoulos, L., A. Karras, and G. A. Krimpas. "Optimizing Apache Spark MLlib: Predictive Performance of Large-Scale Models for Big Data Analytics." *Algorithms*, vol. 18, no. 2, 2025, p. 74. DOI:10.3390/a18020074.
- [6] Salhi, A., et al. "Optimizing High Dimensional Data Classification with a Hybrid AI Driven Feature Selection Framework and Machine Learning Schema." *Scientific Reports*, vol. 15, 2025, art. 35038. DOI:10.1038/s41598-025-08699-4.
- [7] Wang, F., et al. "Navigating the Microarray Landscape: A Comprehensive Review of Feature Selection Techniques and Their Applications." *Frontiers in Big Data*, vol. 8, 2025, art. 1624507. DOI:10.3389/fdata.2025.1624507.
- [8] R. D., and S. M. T. A. "Enhancing Big Data Classification Accuracy Through Deep Learning Techniques." *Artificial Intelligence and Applications*, 2025. DOI:10.47852/bonviewAIA52023432.

- [9] Priyadharsini, M., et al. "Hybrid Synthetic Minority Over-sampling Technique (HSMOTE) and Ensemble Deep Dynamic Classifier Model (EDDCM) for Big Data Analytics." *Scientific Reports*, vol. 15, 2025, art. 39495.
- [10] Palanisamy, S., et al. "Towards Precision in IoT-Based Healthcare Systems: A Hybrid Optimized Framework for Big Data Classification." *Journal of Big Data*, vol. 12, 2025, art. 190.
- [11] Albattah, W., and R. U. Khan. "Impact of Imbalanced Features on Large Datasets." *Frontiers in Big Data*, vol. 8, 2025, art. 1455442. DOI:10.3389/fdata.2025.1455442.
- [12] Carvalho, M., A. J. Pinho, and S. Brás. "Resampling Approaches to Handle Class Imbalance: A Review from a Data Perspective." *Journal of Big Data*, vol. 12, 2025, art. 71.
- [13] Amer, A. A., S. D. Ravana, and R. A. A. Habeeb. "Effective K-Nearest Neighbor Models for Data Classification Enhancement." *Journal of Big Data*, vol. 12, 2025, art. 86. DOI:10.1186/s40537-025-01137-2.
- [14] Farzadnia, Rohham, Payam Shojaei, and Seyed Hadi Mirghaderi. "Implementation of Circularity in Food Supply Chain Based on Big Data Techniques Using Einstein's Fuzzy Methods." *Scientific Reports*, vol. 15, no. 1, 2025.
- [15] Zhang, Bohan, and Ruoyu Tan. "A Trigonometric Function-Based Big Data Model for Network Information Security Control Mechanisms." *Discover Artificial Intelligence*, vol. 5, no. 1, 2025.
- [16] Wang, Lin. "Urban Heat Island Monitoring and Intelligent Prediction Method Based on Big Data Analysis." *Sustainable Energy Research*, vol. 12, no. 1, 2025.
- [17] Wafa, Abeer A., Mai M. Eldefrawi, and Marwa S. Farhan. "Advancing Multimodal Emotion Recognition in Big Data Through Prompt Engineering and Deep Adaptive Learning." *Journal of Big Data*, vol. 12, no. 1, 2025.
- [18] Kumari, Manisha, et al. "An Optimized IDS Framework for Big Data Environments: Integrating Gravitational Search and SMOTE-IPF Data Balancing for High-Accuracy IDS." *SN Computer Science*, vol. 6, no. 7, 2025.
- [19] Mortezaagha, Pouria, Abhisht Makarand Joshi, and Arya Rahgozar. "Inconsistency Detection in Cancer Data Classification Using Explainable-AI." *BMC Artificial Intelligence*, vol. 1, no. 1, 2025.
- [20] Tacaliti, Edoardo, and Jesus S. Aguilar-Ruiz. "Improving Classification on Imbalanced Genomic Data via KDE-Based Synthetic Sampling." *BioData Mining*, vol. 18, no. 1, 2025.