

Anomaly Detection under Class Imbalance: A Survey of Methods, Evaluation Pitfalls, and Open Challenges

Muhammad Jabir Khan

muhammadjabirkhan@gmail.com

Abdul Wali Khan University Mardan

Research Article

Keywords: Anomaly detection, Class imbalance, Machine learning, Deep representation learning, Evaluation pitfalls, AIoT maintenance

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11087383/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Anomaly Detection under Class Imbalance: A Survey of Methods, Evaluation Pitfalls, and Open Challenges

Muhammad Jabir Khan^{1*}

^{1*}Department of Computer Science, Abdul Wali Khan University,
Mardan, Pakistan.

Corresponding author(s). E-mail(s): muhammadjabirkhan@gmail.com;

Abstract

Anomaly detection in industrial and scientific systems is fundamentally constrained by extreme class imbalance, where target events occur with negligible frequency relative to normal operations. This survey provides a structured synthesis of modern computational methodologies designed to mitigate this joint challenge, categorizing existing research into data-level, algorithm-level, generative, and self-supervised paradigms while contrasting their theoretical foundations. Following a PRISMA-guided protocol, this review evaluates the operational tradeoffs of classical machine learning algorithms alongside complex deep neural architectures across five high-stakes application sectors: financial fraud detection, industrial fault diagnostics, healthcare telemetry, cybersecurity network intrusion, and intelligent video surveillance. Beyond algorithmic categorization, this paper exposes critical empirical and theoretical evaluation pitfalls, demonstrating how inappropriate metric selection, flawed validation protocols, and pre-partition data leakage lead to invalidly optimistic performance assessments in skewed scenarios. The survey further identifies major unresolved challenges, including extreme ratio disparity, streaming concept drift, federated edge heterogeneity, and post-hoc interpretability collapse, outlining concrete research directions to guide future system designs.

Keywords: Anomaly detection, Class imbalance, Machine learning, Deep representation learning, Evaluation pitfalls, AIoT maintenance

1 Introduction

Anomaly detection, which identifies data points that deviate significantly from expected normal behavior, is widely used in decision support systems across multiple high-stakes domains [1]. The capacity to isolate these rare deviations is critical because they typically represent high-impact events that can compromise security, financial stability, or human lives. In financial security, identifying unauthorized activities such as credit card fraud protects consumers and institutions from catastrophic losses, where hybrid deep learning architectures combined with data resampling are frequently deployed to manage extreme transaction skews [2–4]. In industrial engineering, identifying minor defects or machine wear enables proactive predictive maintenance to prevent complete system failures and costly downtime [5–7]. Similarly, in healthcare, medical anomaly detection is vital for locating tumors in radiological scans or identifying irregular cardiac activities in patient telemetry, where delayed detection carries fatal consequences [8]. In the digital realm, identifying malicious behaviors in network traffic is key to stopping cybersecurity intrusion attempts [1, 9], whereas intelligent video surveillance systems rely on abnormal event detection to maintain public safety in urban spaces [8, 10].

Despite their importance, the deployment of anomaly detection algorithms in these environments is impeded by severe class imbalance. By definition, anomalies are rare events that constitute a negligible fraction of the overall data distribution, often occurring in ratios below one in a thousand or even one in ten thousand normal transactions [11]. Standard supervised classification algorithms are mathematically optimized to maximize global objective functions, such as overall accuracy, which naturally biases the models toward the majority class [12]. When applied directly to highly skewed datasets, these algorithms achieve near-perfect overall accuracy by simply classifying every instance as normal, which fails to detect any actual anomalies [13]. Consequently, managing this severe class imbalance is the central shared challenge in creating reliable anomaly detection pipelines.

1.1 Delineation and Comparative Survey Positioning

A critical limitation of the prevailing survey literature is that existing works treat class imbalance and anomaly detection in isolation, failing to examine their theoretical and empirical convergence. As shown in Table 1, foundational surveys on imbalanced learning [12, 14] systematically examine resampling heuristics, cost-sensitive matrices, and ensemble adaptations across standard multiclass benchmarks; however, they inherently assume that the minority class contains sufficient structural density to construct valid interpolations, which breaks down in extreme anomaly scenarios. Conversely, general anomaly detection reviews [15, 16] primarily formalize unsupervised one-class profiling or reconstruction baselines, largely overlooking scenarios where scarce, highly skewed positive labels exist and can be harnessed. Furthermore, domain-focused reviews such as Ren et al. [5] investigate class imbalance strictly within rotating machinery fault diagnostics, without cross-pollinating algorithmic solutions across transactional, clinical, or cybersecurity environments.

Table 1 Comparative Positioning of This Survey Relative to Related Literature

Survey Reference	Extreme Skew Handling	Semi-Supervised Anomaly Scenarios	Methodological Taxonomy	Evaluation Pitfalls & Data Leakage	Cross-Domain Synthesis	Primary Scope & Key Limitations
Chandola et al. [15]	Partial	Unsupervised	Classical ML	None	Moderate	Foundational anomaly taxonomy; lacks deep learning and modern skew mitigation.
Krawczyk [12]	High	Classification	Resampling/Cost	Low	Moderate	Imbalanced learning survey; does not target anomaly detection pipelines.
Pang et al. [16]	Low	Unsupervised	Deep Architectures	Low	High	Deep anomaly detection review; assumes absence of labeled anomalies.
Ren et al. [5]	High	Supervised/One-Class	Hybrid Networks	Moderate	Domain-Specific	Restricted strictly to industrial rotating machinery and bearing fault diagnostics.
Chen et al. [14]	High	Classification	Modern Ensembles	Low	High	Comprehensive imbalanced learning review; lacks explicit anomaly-specific loss analysis.
This Survey	High	Unified (Un/Semi/Super)	Comprehensive	Extensive	Cross-Domain (5)	Unifies extreme skew, modern self-supervised deep models, leakages, and AIoT.

This survey explicitly bridges this divide. Unlike Chen et al. [14], who examine imbalanced learning at large, we address the specific operational boundary where class ratios exceed 100:1 to 10,000:1, and where traditional oversampling techniques induce catastrophic class overlap or temporal contamination. In contrast to Ren et al. [5], who specialize purely in industrial sensors, our investigation unifies five distinct high-stakes domains while diagnosing widespread empirical pitfalls (e.g., pre-partition resampling leakage and metric distortions).

The primary contributions of this survey are:

1. **Unified Multidimensional Taxonomy:** We establish a structured taxonomy organizing data-level, algorithm-level, classical machine learning, deep representation, generative, and self-supervised paradigms specifically designed for imbalanced anomaly detection.
2. **Cross-Domain Operational Synthesis:** We synthesize empirical characteristics, benchmark datasets, and domain constraints across five critical sectors: financial fraud detection, industrial fault diagnosis, healthcare monitoring, network cybersecurity, and video surveillance.
3. **Systematic Evaluation and Pitfall Audit:** We mathematically dissect why standard metrics fail under extreme data skews, demonstrate how flawed experimental protocols (e.g., pre-split oversampling, temporal contamination) yield invalidly optimistic results, and propose standardized validation protocols.
4. **Actionable Research Roadmap:** We categorize and articulate pressing open challenges, including real-time drifting stream adaptation, privacy-preserving federated edge learning with non-IID skews, explainability stability under minority sparsity, and LLM integration.

2 Survey Methodology

To ensure a comprehensive, representative, and unbiased synthesis of the existing literature at the intersection of anomaly detection and class imbalance, we established a structured literature search protocol modeled after the PRISMA 2020 guidelines [17].

The literature gathering process was executed across three primary digital databases: Google Scholar, IEEE Xplore, and ACM Digital Library.

2.1 Search Strategy and Query Formulation

We restricted our primary search scope to peer-reviewed journal publications and top-tier conference proceedings published between January 2014 and June 2026, supplemented with seminal historical papers (e.g., [18–21]). The search queries merged core thematic concepts using Boolean logic:

1. ("anomaly detection" OR "outlier detection" OR "one-class classification")
2. AND ("class imbalance" OR "imbalanced learning" OR "minority oversampling" OR "skewed distribution")
3. AND ("evaluation metrics" OR "data leakage" OR "concept drift" OR "prognostics")

These core terms were modified recursively to target specific domain applications, such as "credit card fraud", "vibration diagnostics", "AIoT maintenance", "sepsis prediction", "network intrusion", or "video surveillance".

2.2 Inclusion and Exclusion Criteria

To guarantee a high-quality, reproducible, and balanced bibliography, we filtered retrieved articles using strict, pre-defined eligibility rules:

- **Inclusion Criteria:** (1) Original research papers proposing or evaluating data-level, algorithm-level, deep architectural, or self-supervised solutions for imbalanced anomaly detection; (2) study designs demonstrating quantitative evaluation using standard, rigorous metrics (e.g., MCC, AUC-PR, F1-score, Precision-Recall); (3) studies published in peer-reviewed journals or top-tier conference proceedings written in English.
- **Exclusion Criteria:** (1) Papers addressing generic balanced classification without severe class skews; (2) anomaly detection studies assuming perfectly clean or balanced training data; (3) informal blog posts, white papers, or commercial reports lacking peer review; (4) duplicate publications or highly incremental updates of existing algorithms.

As formalized in Figure 1, the database search initially identified 236 records. After removing 68 duplicate entries, 168 records underwent title and abstract screening, resulting in the exclusion of 64 records. A total of 104 full-text articles were rigorously assessed against our eligibility criteria, from which 20 were excluded due to lack of metric rigor or missing methodological detail. The resulting core synthesis encompasses 84 primary peer-reviewed studies and foundational benchmarks.

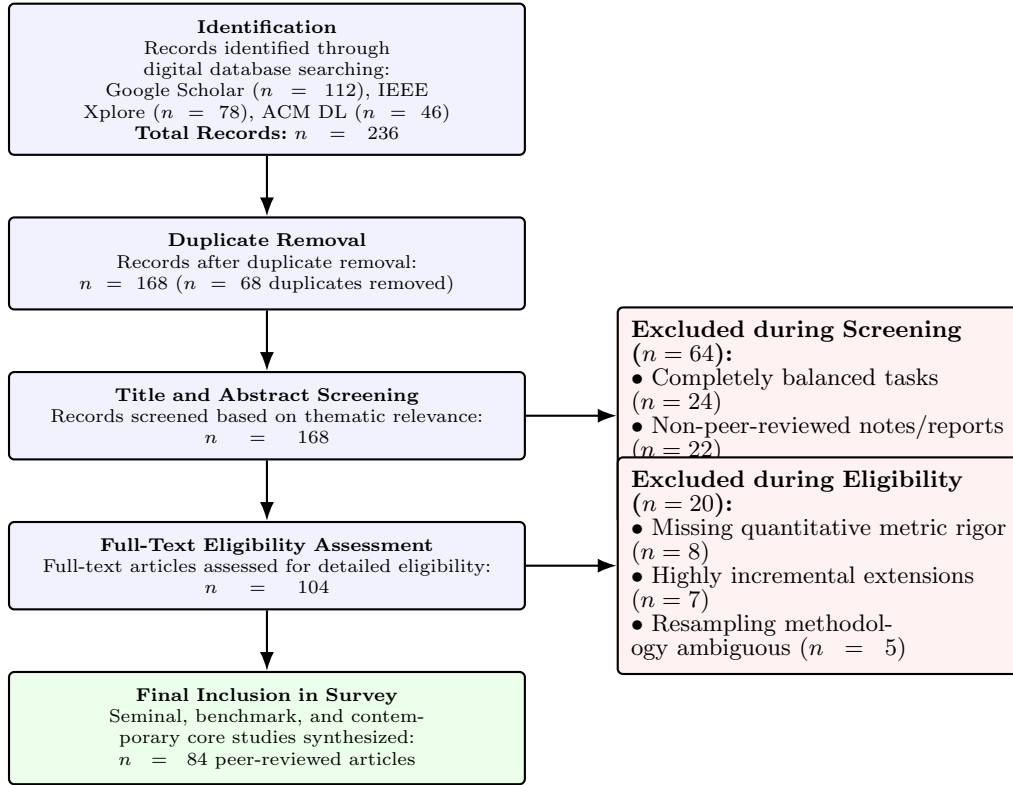


Fig. 1 PRISMA-guided literature identification, screening, exclusion, and inclusion flowchart.

2.3 Bibliometric and Coverage Distribution

To illustrate the temporal evolution and domain balance of the analyzed literature, Figure 2 depicts the publication timeline and domain distribution of the reviewed studies. The synthesis demonstrates steady expansion in deep anomaly detection under class imbalance since 2017, culminating in significant recent activity (2023–2026) driven by self-supervised and industrial AIoT architectures.

3 Background and Definitions

3.1 Anomaly Detection

The formal categorization of anomalous behaviors is fundamental to selecting appropriate detection algorithms. According to established machine learning taxonomies, anomalies are broadly classified into point, contextual, and collective anomalies [15]. A point anomaly occurs when an individual data instance is considered anomalous with respect to the remainder of the dataset [15]. For instance, in financial transactions, a single credit card transaction of ten thousand dollars on an account that historically experiences a maximum daily charge of fifty dollars represents a classic point

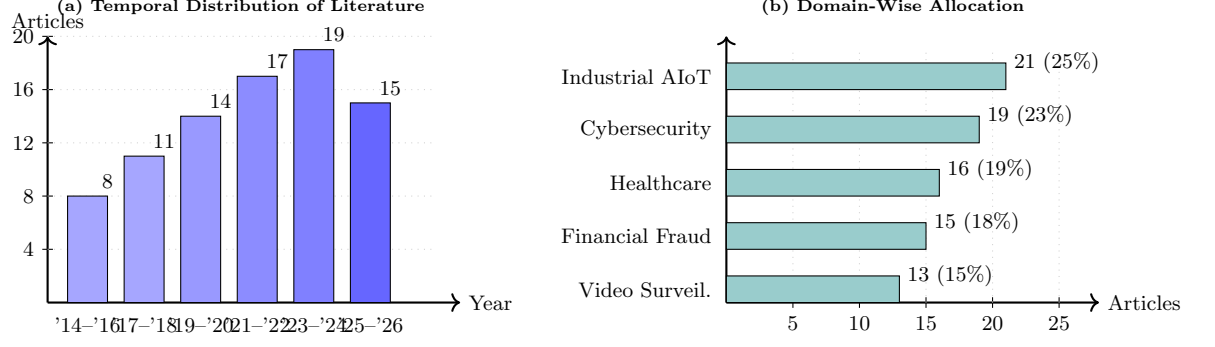


Fig. 2 Bibliometric breakdown of surveyed research: (a) publication volume timeline across 2014–2026; (b) distribution of reviewed papers across five target application sectors.

anomaly. In contrast, a contextual anomaly occurs when a data instance is anomalous in a specific context but normal in another [15]. For example, in healthcare telemetry, a patient body temperature of thirty-nine degrees Celsius is entirely normal during intense physical exercise but represents a contextual anomaly if recorded while the patient is sleeping. Finally, a collective anomaly occurs when a collection of related data instances is anomalous with respect to the entire dataset, even if the individual instances are not anomalous when analyzed in isolation [15]. In cybersecurity, a continuous, rapid sequence of tiny, low-byte ping packets from a single internet protocol address represents a collective anomaly indicating a slow-rate denial of service attack, whereas any single ping packet on its own is entirely benign. Understanding these anomaly types is crucial, as their occurrence is intimately linked to the highly skewed distribution of real-world data.

3.2 Class Imbalance

In machine learning, class imbalance occurs when the joint probability distribution of features and class labels is characterized by a significant disparity in class frequencies [11]. Under a binary classification framework, this problem is defined by the coexistence of a majority class, which contains the overrepresented normal data instances, and a minority class, which contains the underrepresented anomalous target instances [11]. To quantify this disparity, researchers utilize the imbalance ratio (IR), which is mathematically defined as the cardinality of the majority class divided by the cardinality of the minority class:

$$\text{IR} = \frac{|\mathcal{D}_{\text{majority}}|}{|\mathcal{D}_{\text{minority}}|} = \frac{N_-}{N_+} \quad (1)$$

Unlike artificial imbalances caused by faulty experimental sampling, class imbalance in anomaly detection is a structural, natural property of the monitored environment. For example, in industrial quality control, automated assembly lines manufacture millions of high-quality components daily, meaning machine faults and manufacturing defects are rare, structural outliers. Thus, the extreme skew in the imbalance ratio is

an inherent feature of the physical process being monitored rather than an experimental error. This inherent ratio disparity directly undermines the efficacy of standard evaluation frameworks.

3.3 Why Standard Metrics Fail

Standard classification metrics, most notably overall classification accuracy, provide highly misleading performance estimates when applied to imbalanced datasets. In a scenario where the imbalance ratio is 999:1, a trivial model that classifies all instances as normal achieves a classification accuracy of 99.9% [11]. This overall accuracy is misleading because the model fails to detect the single anomaly it was deployed to identify, rendering the system useless in practice. To address this failure, robust alternative metrics must be employed. Precision measures the ratio of true positive detections to all predicted positive instances, capturing the model resistance to false alarms:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Recall, also known as sensitivity, measures the proportion of actual anomalies correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

The F1-score is defined as the harmonic mean of precision and recall, balancing both metrics without letting one skew the result [22]:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) explores the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across all classification thresholds. However, because $\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$, when the normal class (TN) is colossal, thousands of false alarms can accumulate while FPR remains deceptively negligible [23]. Conversely, the Area Under the Precision-Recall Curve (AUC-PR) evaluates precision against recall, isolating minority-class behavior without incorporating TN, making it substantially more sensitive to false positives under extreme skews.

To evaluate performance balance on both classes simultaneously, the geometric mean computes the square root of the product of recall and specificity [11]:

$$\text{G-Mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \cdot \frac{\text{TN}}{\text{TN} + \text{FP}}} \quad (5)$$

Finally, the Matthews Correlation Coefficient (MCC) provides an invariant, symmetric evaluation of binary classification performance even when classes are of extremely

disparate sizes by taking into account all four quadrants of the confusion matrix [24]:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (6)$$

The Matthews correlation coefficient yields a score between -1 and $+1$, where $+1$ represents a perfect prediction, 0 represents performance no better than random guessing, and -1 indicates total disagreement [24].

3.4 The Intersection: Optimization Bias and Margin Collapse

The intersection of class imbalance and anomaly detection creates a severe learning bias during model optimization rather than just performance degradation during evaluation. Standard loss functions, such as cross-entropy, treat all data samples equally during the optimization process, which causes the gradient of the loss function to be dominated by the majority class [25]:

$$\nabla_{\theta} \mathcal{L}(\theta) = \frac{1}{N} \sum_{i \in \mathcal{D}_-} \nabla_{\theta} \ell(f(x_i; \theta), y_i) + \frac{1}{N} \sum_{j \in \mathcal{D}_+} \nabla_{\theta} \ell(f(x_j; \theta), y_j) \quad (7)$$

When $N_- \gg N_+$, the first summation dominates the gradient updates by several orders of magnitude. Consequently, gradient descent updates parameters along trajectories that reduce majority class residual error, effectively ignoring minority anomaly representations [13, 25]. This gradient dominance causes a severe decision boundary shift in the feature space, compressing the geometric margin of the minority class and leading to high false negative rates.

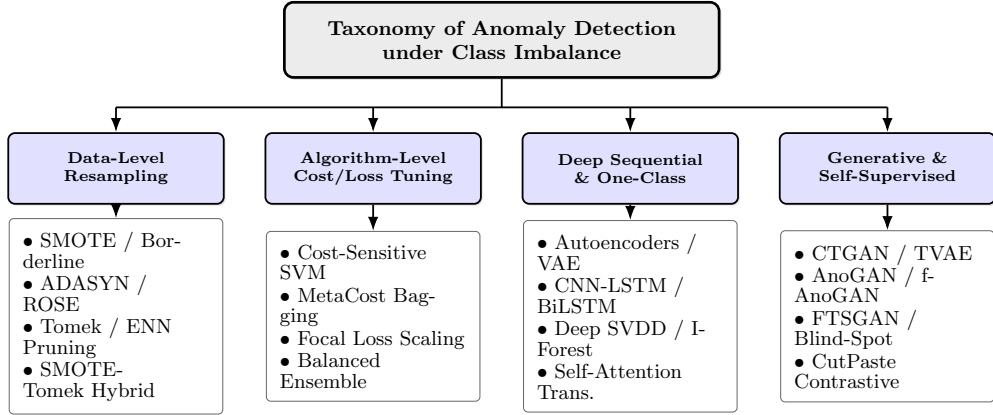


Fig. 3 Comprehensive taxonomy of anomaly detection paradigms under extreme class imbalance.

4 Taxonomy of Approaches

4.1 Data-Level Methods

Data-level methods seek to rebalance the training distribution prior to model optimization, thereby mitigating gradient dominance by the majority class. Among oversampling techniques, the Synthetic Minority Over-sampling Technique (SMOTE) is the foundational framework [21]. SMOTE operates by selecting a minority class instance x and identifying its k -nearest neighbors in the feature space. A neighbor x_{nn} is chosen at random, and a synthetic sample x_{new} is generated via linear interpolation:

$$x_{new} = x + \lambda(x_{nn} - x) \quad (8)$$

where x represents a randomly selected minority class sample, x_{nn} is one of its k -nearest neighbors in the minority class feature space, and λ represents a uniform random variable in the interval $[0, 1]$ [21]. Because basic SMOTE interpolates indiscriminately, it can inadvertently generate synthetic instances inside majority class regions, exacerbating class overlap. To address this, Borderline-SMOTE refines the selection process by categorizing minority instances into safe, danger, or noise based on their local neighborhood [26]. Synthetic generation is restricted strictly to the danger instances, which lie near the decision boundary [26]. Support Vector Machine SMOTE (SVMSMOTE) uses a support vector machine classifier to locate the optimal decision boundary and prioritizes oversampling along the support vectors, thereby reinforcing boundary definition [27]. The Adaptive Synthetic (ADASYN) sampling approach dynamically scales the number of synthetic samples generated for each minority instance based on its difficulty of learning, assigning higher generation rates to minority instances surrounded by majority class neighbors [28]. Alternatively, Random Over-Sampling Examples (ROSE) adopts a smoothed bootstrap approach, generating synthetic data points for both classes from conditional kernel density estimates, which avoids the collinearity and overfitting pitfalls associated with strict linear interpolation [29].

Undersampling methods, conversely, reduce the cardinality of the majority class to balance the class ratio. While random undersampling discards majority samples indiscriminately at the risk of losing critical discriminative information, heuristic approaches prune data strategically. Tomek Links identify pairs of nearest neighbors from opposite classes that lie in close proximity [19]. Removing the majority instance of the Tomek Link, or both instances, effectively cleans the decision boundary and broadens the separation margin [19]. Similarly, the Edited Nearest Neighbours (ENN) rule evaluates each sample using its k -nearest neighbors and discards any instance whose class label differs from the majority of its neighbors, smoothing the decision boundary by eliminating noisy, overlapping instances [18]. The NearMiss family offers three structured selection strategies based on k -nearest neighbor distances: NearMiss-1 selects majority class samples with the minimum average distance to the closest minority class samples; NearMiss-2 selects majority samples with the minimum average distance to the farthest minority class samples; and NearMiss-3 selects a guaranteed number of majority samples closest to each minority instance to preserve boundary

details [30]. To minimize information loss, the ClusterCentroids method groups majority class instances using K -means clustering and replaces the original samples within each cluster with their computed cluster centroid, reducing majority class cardinality while maintaining the global distribution profile [31].

Hybrid resampling methods combine oversampling and undersampling to exploit their complementary strengths while mitigating their individual limitations. Oversampling techniques populate sparse regions but often create artificial class overlap, whereas undersampling cleans noise but can discard valuable data. Combining these approaches, such as in SMOTE+Tomek and SMOTE+ENN, first applies oversampling to bolster minority class representation and then utilizes undersampling heuristics to prune noisy, overlapping synthetic and real instances from the decision boundary [2]. Notably, hybrid configurations combining SMOTE-Tomek resampling with sequential deep architectures like BiLSTM and GRU have been deployed to address transaction imbalances in credit card fraud detection [4].

More recently, deep generative architectures have emerged to model highly complex tabular distributions. The Conditional Tabular GAN (CTGAN) addresses the challenges of mixed-type tabular data by employing mode-specific normalization for continuous columns and training-by-sampling with a conditional generator to ensure rare categorical categories are sufficiently represented during training [32]. The Tabular Variational Autoencoder (TVAE) provides a non-adversarial alternative, utilizing variational inference to map input features to a continuous latent space by optimizing the evidence lower bound [32]. This optimization avoids the training instability and mode collapse common to generative adversarial networks while yielding high-fidelity synthetic samples [32]. Conditional generative adversarial networks integrate class labels or auxiliary conditioning vectors to guide the generator, allowing for targeted synthesis of minority anomalies to resolve severe class skews [33].

4.2 Algorithm-Level Methods

Whereas data-level methods alter the underlying training distribution, algorithm-level methods modify the inner optimization mechanics of classification models to natively tolerate extreme class imbalances. A primary paradigm within this category is cost-sensitive learning, which incorporates asymmetric misclassification costs directly into the model objective function. Instead of treating all errors equally, cost-sensitive learning assigns a significantly higher penalty to false negatives than to false positives, aligning the optimization process with the practical reality that failing to detect a critical anomaly is far more damaging than triggering a false alarm [34]. For instance, a cost-sensitive Support Vector Machine (CS-SVM) modifies the standard soft-margin formulation as defined in Equation 9 [35]:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C^+ \sum_{i:y_i=+1} \xi_i + C^- \sum_{i:y_i=-1} \xi_i \quad (9)$$

where w and b define the separating hyperplane, $\xi_i \geq 0$ are the slack variables measuring margin violations, and C^+ and C^- represent the asymmetric cost penalty

parameters assigned to the positive (minority) and negative (majority) classes, respectively. When training on highly imbalanced environments, setting $C^+ \gg C^-$ forces the optimizer to favor minority-class decision boundary alignment. At the meta-algorithmic level, MetaCost serves as a general wrapper method designed to make any arbitrary classifier cost-sensitive [36]. MetaCost utilizes bootstrap bagging to estimate the conditional class probabilities for each training instance and then systematically relabels the training dataset according to the class that minimizes the expected misclassification cost [36].

Alternatively, Focal Loss (FL) has emerged as an elegant algorithm-level mechanism to natively handle extreme skews by dynamically scaling standard cross-entropy loss based on prediction confidence [37]:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (10)$$

where $p_t \in [0, 1]$ represents the model’s estimated probability for the ground-truth class, $\alpha_t \in [0, 1]$ is an asymmetric addressing parameter for overall imbalance, and $\gamma \geq 0$ is a focusing parameter. The term $(1 - p_t)^\gamma$ automatically down-weights the loss contributions of easily classified majority samples, focusing the optimization gradients strictly on rare, hard-to-classify anomalies.

An alternative algorithm-level strategy is one-class classification, which reformulates the learning objective entirely by profiling normality rather than attempting to separate highly imbalanced classes. One-Class Support Vector Machines (OCSVM) map training data into a high-dimensional feature space using kernel functions and optimize a hyperplane to separate the majority of normal instances from the coordinate origin with a maximum margin [20]. This formulation is controlled by a parameter that determines the upper bound on the fraction of training errors and the lower bound on the number of support vectors [20]. Support Vector Data Description (SVDD) adopts a similar kernel-based approach but optimizes a hyper-spherical boundary around the normal data instances rather than a hyperplane [38]. Instances that fall outside the radius of this hyper-sphere are classified as anomalies [38]. Moving away from distance-based boundary profiling, the Isolation Forest algorithm isolates anomalous instances based on their structural susceptibility to recursive partitioning [39]. The anomaly score $s(x, n)$ for an instance x given a sample size of n is defined as:

$$s(x, n) = 2^{-\frac{\mathbb{E}(h(x))}{c(n)}} \quad (11)$$

where $h(x)$ is the path length of instance x in a randomized decision tree, $\mathbb{E}(h(x))$ is the expected path length across an ensemble of trees, and $c(n) = 2 \ln(n - 1) + 0.5772156649 - \frac{2(n-1)}{n}$ represents the average path length of an unsuccessful search in a Binary Search Tree (BST) built from n instances, acting as a normalization factor. An anomaly score $s \rightarrow 1$ indicates highly isolated instances (anomalies) because they possess unique feature characteristics and are highly sparse, requiring significantly fewer random splits to isolate within a randomized decision tree, manifesting as short path lengths from the root node [39].

Ensemble-based algorithm-level methods combine multiple weak learners to improve generalization, leveraging ensemble diversity to mitigate the bias caused by class imbalance. The `BalancedBaggingClassifier` constructs bootstrap samples and applies random undersampling to the majority class within each bootstrap bag prior to training each base estimator, preserving data diversity while ensuring balanced training sub-problems [31]. To utilize the discarded majority class data more effectively, `EasyEnsemble` independently draws multiple undersampled subsets of the majority class, trains an AdaBoost ensemble on each subset alongside all minority class instances, and combines their collective predictions to produce a highly robust consensus decision [40]. `BalanceCascade` extends this concept into a sequential learning paradigm, training classifiers iteratively and discarding majority class instances that are correctly classified with high confidence at each step, thereby systematically narrowing the optimization focus on ambiguous boundary instances [40]. Alternatively, `RUSBoost` integrates random undersampling directly into the boosting framework [41]. At each boosting step, `RUSBoost` undersamples the majority class of the weighted training distribution, allowing the AdaBoost update rules to focus on difficult, misclassified minority and majority instances without suffering from the extreme gradient dominance of the majority class [41].

4.3 Classical Machine Learning under Imbalance

Classical machine learning algorithms, when applied naively to highly skewed distributions, exhibit structural biases that compromise anomaly detection efficacy. Logistic regression, which optimizes a log-likelihood function to maximize global accuracy, naturally shifts its decision boundary intercept toward the majority class under severe imbalance [42]. This intercept shift mathematically compresses the predicted probabilities for the minority class, causing the model to underestimate the likelihood of rare events, which induces an excessively high false negative rate [42]. To correct this, practitioners must manually adjust the classification threshold away from the default midpoint, though this intervention often degrades overall precision. Similarly, Support Vector Machines suffer from decision boundary skew. Because the standard soft-margin formulation relies on a penalty parameter that treats all misclassifications uniformly, the sheer density of majority instances forces the optimal separating hyperplane aggressively toward the minority class space to minimize total margin violations [43]. This skewed boundary severely erodes the geometric decision margin for anomalies.

Tree-based ensemble methods exhibit relatively more robustness to skewed data but remain fundamentally majority-biased. Random Forest constructs an ensemble of unpruned decision trees using bootstrap aggregation and random feature selection, injecting structural diversity into the learning process [44]. Despite this robustness, standard splitting criteria, such as Gini impurity or information gain, evaluate the probability of class labels at each node. Because majority class instances dominate the node distributions, the impurity reduction overwhelmingly favors splits that isolate majority patterns, leaving minority instances misclassified in impure terminal leaves [45]. Extreme Gradient Boosting optimizes via sequential gradient descent, fitting subsequent trees to the residuals of previous ones. This boosting mechanism is highly

sensitive to class imbalance because the gradient updates are fundamentally dominated by majority class errors [46].

To counteract these mathematical biases natively, standard implementations utilize class weight parameters. Random Forest uses class weights to dynamically adjust the Gini penalty during node splitting, while Extreme Gradient Boosting uses the `scale_pos_weight` parameter to directly scale the gradient of the loss function for positive, minority instances [46]. This hyperparameter tuning mathematically simulates a balanced distribution without requiring data-level resampling, providing a robust baseline before transitioning to more complex deep learning architectures.

4.4 Deep Learning Architectures

Deep learning architectures offer representation learning capabilities for complex, high-dimensional data, yet they require specific structural adaptations to perform anomaly detection under severe class imbalance. Unsupervised deep learning frameworks circumvent the class imbalance problem entirely by profiling only the majority normal class. Standard autoencoders achieve this by employing an encoder network that compresses normal input data into a lower-dimensional latent bottleneck, followed by a decoder network that reconstructs the original input [16]. The network is trained exclusively on majority class data to minimize the mean squared error of reconstruction. Consequently, when anomalous instances are passed through the network, the model fails to reconstruct their unfamiliar feature patterns, yielding a reconstruction error that serves as an anomaly score [16]. However, deterministic autoencoders can reconstruct anomalies if the bottleneck is too wide, leading to elevated false negative rates.

Variational Autoencoders address this limitation by regularizing the latent space probabilistically. Instead of mapping inputs to fixed deterministic vectors, the encoder of a Variational Autoencoder maps inputs to a probability distribution defined by a mean and a variance vector [47]. To permit gradient descent backpropagation through this stochastic sampling process, the model utilizes the reparameterization trick [47]. The network optimizes the Evidence Lower Bound (ELBO) loss function, which precisely balances reconstruction fidelity against a Kullback-Leibler divergence penalty:

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; x) = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - D_{\text{KL}}(q_{\phi}(z|x) \parallel p(z)) \quad (12)$$

where $q_{\phi}(z|x)$ represents the variational approximate posterior (encoder), $p_{\theta}(x|z)$ is the generative likelihood (decoder), $p(z)$ is the prior (typically a standard Gaussian $\mathcal{N}(0, I)$), and D_{KL} represents the Kullback-Leibler divergence measuring the statistical distance between the approximate posterior and prior distributions. This divergence forces the learned latent distribution to match a standard normal prior, ensuring a continuous and densely packed latent space. By minimizing latent space distortion, Variational Autoencoders robustly measure the statistical probability of an instance, identifying out-of-distribution anomalies with greater rigor than threshold-based deterministic reconstruction.

For sequential anomaly detection, recurrent neural networks are deployed to manage temporal dependencies, though basic architectures suffer from vanishing gradients. Long Short-Term Memory networks solve this by utilizing a distinct cell state regulated

by input, forget, and output gates, which selectively retain or discard sequential information over extended time steps [48]. Gated Recurrent Units simplify this paradigm by merging the cell state and hidden state, utilizing only an update gate and a reset gate [3]. This structural consolidation makes Gated Recurrent Units computationally lighter while maintaining comparable temporal memory capabilities. To capture context from both past and future states simultaneously, bidirectional implementations, such as BiLSTM and BiGRU, process the input sequence through parallel forward and backward hidden layers.

Recent developments in industrial Artificial Intelligence of Things (AIoT) have extensively leveraged these hybrid spatio-temporal architectures to resolve severe sensor imbalance and temporal noise. For instance, Nguyen-Da et al. [6] established a real-time AIoT anomaly detection framework for industrial diesel generators using a coordinated CNN-LSTM architecture that efficiently decouples dynamic rotational fluctuations from catastrophic operational failures under Industry 4.0 conditions. Similarly, Nguyen-Thanh [49] demonstrated that fusing Improved Particle Swarm Optimization (IPSO) with hybrid BiLSTM-CNN pipelines allows reliable fault isolation in municipal firefighting pump telemetry despite extreme historical sparsity of failure states. To achieve online classification across continuous multi-sensor regimes, Nguyen-Thanh and Cho [7] and Nguyen et al. [50] formulated two-stage prognostic deep networks integrating CNN-BiLSTM blocks, successfully demonstrating that structural temporal gating mitigates majority-class gradient bias in streaming vibration data.

In domains involving spatial or grid-like data, Convolutional Neural Networks employ localized convolutional kernels and pooling layers to extract hierarchical spatial representations [51]. While traditional two-dimensional convolutions excel in image-based anomaly detection, one-dimensional Convolutional Neural Networks are effective at capturing localized signal motifs in time-series telemetry or industrial vibration data. To maximize performance on multidimensional tasks, single architectures are superseded by hybrid models. A CNN-LSTM or CNN-GRU hybrid architecture systematically partitions the learning process. The convolutional layers act as spatial feature extractors, compressing the data into dense feature vectors that are sequentially fed into the recurrent layers to model temporal dynamics. These hybrids consistently outperform standalone architectures by fusing local structural extraction with deep temporal modeling, addressing both spatial and temporal anomaly components.

Recently, Transformer architectures have transformed sequence-based anomaly detection by completely discarding recurrence in favor of self-attention mechanisms [52]. Because Transformers lack sequential step-by-step processing, they utilize positional encodings to retain the temporal order of the data. The core multi-head self-attention mechanism computes attention weights by calculating the scaled dot product of query, key, and value matrices, allowing the model to weigh the importance of every element in a sequence relative to all other elements simultaneously [52]. Recurrent networks naturally degrade in performance when sequences exceed a few hundred steps, whereas Transformers maintain direct computational connections between all

sequence elements regardless of their distance [52]. This parallel processing capability allows Transformers to capture long-range dependencies across extended temporal horizons without sequential bottlenecks. Consequently, self-attention models are adept at detecting stealthy, low-and-slow anomalies in complex environments.

4.5 Generative, Self-Supervised, and Advanced Representation Paradigms

Generative and self-supervised approaches represent a paradigm shift in addressing the extreme scarcity of labeled anomalies by leveraging unsupervised representation learning and artificial pretext tasks to model the boundary of normality. Generative Adversarial Networks (GANs) have pioneered this area, primarily through the AnoGAN framework [53]. AnoGAN is trained exclusively on normal data, allowing the generator to capture the manifold of normal anatomical or structural variability. During inference, detecting an anomaly requires mapping a query image back into the latent space of the generator [53]. This mapping is achieved through an iterative gradient descent optimization that searches for a latent code z that minimizes a dual objective function comprising a residual reconstruction loss in the image space and a discrimination loss in the feature space of the discriminator [53]. The cumulative loss at the optimal latent state serves as the final anomaly score. To overcome the high computational latency of this iterative search, f-AnoGAN introduces an encoder network trained on the GAN generator using a Wasserstein GAN objective [54]. The encoder maps query images directly to the latent space in a single feed-forward pass, accelerating inference speeds and making real-time anomaly detection viable [54].

Beyond standard spatial GANs, novel spectral-frequency mappings have emerged to resolve boundary ambiguity under severe imbalance. For instance, Wang et al. [55] introduced the Frequency-to-Spectrum Mapping GAN (FTSGAN) for semi-supervised hyperspectral anomaly detection. By mapping representations into the fractional Fourier domain, FTSGAN synthesizes high-fidelity spectral transitions that prevent generative networks from collapsing into majority-background modes.

Concurrently, blind-spot and blind-block architectures have resolved the longstanding “identity-mapping” pathology of standard autoencoders, where deep networks inadvertently learn to reconstruct anomalous regions with high fidelity. The Blind-Block Reconstruction Network with a Guard Window (BockNet) [56] enforces a spatial guard window over the receptive field center, preventing the network from accessing target pixels during reconstruction and thereby amplifying the reconstruction error strictly for sparse anomalous targets. Extending this concept, Wang et al. [57] developed the Global Feature-Injected Blind-Spot Network (PUNNet), combining patch-shuffle downsampling with nonlinear activation-free modules to inject global contextual constraints into the blind-spot reconstruction pipeline, effectively preventing over-reconstruction of rare anomalies.

In graph-structured environments, where extreme topological imbalance complicates node classification, Huang et al. [58] formulated a cross-domain graph anomaly detection framework combining graph transfer with graph decouple mechanisms. By decoupling domain-invariant structural attributes from localized edge connectivity

skews, this approach prevents the dense normal class of the source domain from overwhelming the target domain feature alignment.

In scenarios where limited anomalous samples are available, conditional generative adversarial networks (cGANs) are utilized to explicitly model the minority distribution. By conditioning both the generator and discriminator on class labels, cGANs synthesize realistic, high-fidelity minority class instances, allowing targeted oversampling of scarce anomaly modes [59]. This approach is complemented by contrastive learning frameworks, which learn robust representations of normal data without relying on generative reconstruction. Contrastive techniques, such as the Contrastive Shifted Instances (CSI) framework, optimize a contrastive loss that pulls augmented views of the same normal sample together in the latent space while pushing distributionally shifted versions apart [60]. This optimization maps normal data into a highly compact, hyperspherical feature cluster, causing out-of-distribution anomalies to project far outside this boundary during testing [60].

Self-supervised models extend this representation learning capability by training deep networks on artificial pretext tasks designed to expose anomalous structures. In image and signal domains, predicting geometric transformations, such as rotation, translation, and shearing, serves as an effective pretext task [61]. The network learns to classify which transformation was applied to a normal input; during testing, the model’s prediction confidence collapses when presented with an anomaly, as the anomalous features do not conform to the learned transformation rules [61]. Similarly, the CutPaste framework creates self-supervised representations for industrial inspection by cutting local patches from a normal image and pasting them at random coordinates to simulate defects [62]. Training a binary classifier to distinguish between normal images and these artificial CutPaste augmentations forces the feature extractor to capture localized, high-resolution structural anomalies [62].

4.6 Comparison of Methodological Paradigms

A comprehensive comparison of the discussed anomaly detection and class imbalance methods is presented in Tables 2 and 3. The tables compare data-level, algorithm-level, classical machine learning, deep learning, generative, and self-supervised paradigms across their core mechanisms, computational overhead, and unique operational limitations when handling severe class skews.

5 Domain-Specific Applications

5.1 Financial Fraud Detection

Anomaly detection in financial systems is critical due to the staggering economic losses generated by malicious activities such as credit card fraud, money laundering, and identity theft. In this domain, class imbalance is exceptionally severe because fraudulent transactions represent an incredibly tiny fraction of global daily financial activities. For example, in the widely used Kaggle Credit Card Fraud dataset, which comprises transactions made by European cardholders, there are only 492 confirmed fraudulent cases out of 284,807 total transactions, yielding an extreme imbalance ratio

Table 2 Systematic Comparison of Anomaly Detection and Class Imbalance Methods (Part 1)

Method	Category	Core Mechanism	Primary Domain	Handles Imbalance How	Computational Cost	Key Limitation
SMOTE	Data-Level	Synthesizes minority instances by linear interpolation along lines joining neighbors.	Tabular	Artificially inflates minority class count to balance the dataset.	Low	Can create synthetic instances inside majority class regions, causing class overlap.
ADASYN	Data-Level	Generates synthetic instances adaptively based on minority density distribution.	Tabular	Direct oversampling focusing heavily on hard-to-learn boundary regions.	Low to Medium	High sensitivity to outliers and noise present within the minority class.
Tomek Links	Data-Level	Finds pairs of closest instances of opposite classes and removes the majority member.	Tabular	Direct under-sampling to widen the decision margin and clean overlap.	Medium	May not sufficiently reduce majority class cardinality in extremely skewed datasets.
SMOTE+Tomek	Data-Level	Sequentially applies SMOTE oversampling followed by Tomek Link majority pruning.	Tabular	Simultaneously increases minority representation while cleaning class boundary overlaps.	Medium	Higher computational latency than single techniques due to sequential operations.
CTGAN	Data-Level	Uses conditional generator and training-by-sampling on tabular distributions.	Tabular	Generates tabular samples based on conditional distributions.	High	Suffers from training instability and requires substantial hyperparameter tuning.
Cost-sensitive SVM	Algorithm-Level	Imposes asymmetric misclassification penalties on classes in SVM loss function.	Tabular	Natively penalizes false negatives more severely than false positives.	Medium to High	Quadratic optimization complexity; scales poorly to massive streaming datasets.
Isolation Forest	One-Class	Isolates anomalies by recursively partitioning random features using decision	Tabular	Exploits minority sparsity because anomalies require fewer partitions to isolate.	Low	Struggles to capture highly correlated non-linear feature relationships.

Table 3 Systematic Comparison of Anomaly Detection and Class Imbalance Methods (Part 2 – Continued)

Method	Category	Core Mechanism	Primary Domain	Handles Imbalance How	Computational Cost	Key Limitation
XGBoost	Classical ML	Iteratively trains gradient boosted trees focusing sequentially on residuals.	Tabular	Scales gradient weights of minority instances using parameter tuning.	Medium	Highly sensitive to hyperparameter choices under extreme class skew.
Autoencoder	Deep Learning	Compresses input features to a narrow bottleneck and reconstructs them.	Images, Tabular	Profiles normality; high reconstruction error indicates out-of-distribution anomaly.	Medium to High	Wide bottleneck can lead to perfect reconstruction of anomalous samples.
VAE	Deep Learning	Encodes inputs to latent distribution parameters and regularizes with KL divergence.	Images, Tabular	Evaluates sample probability; out-of-distribution instances yield poor probability.	Medium to High	Complex optimization is required to balance reconstruction and latent regularization.
LSTM / BiLSTM	Deep Learning	Uses gating mechanisms to regulate information flow across temporal sequences.	Sequences	Models sequential normal behavior; deviations in output predict anomalies.	High	Training is sequentially bound and cannot easily scale across long-range histories.
CNN-BiLSTM / AIoT Hybrids	Deep Learning	Fuses localized convolution with bidirectional gating for industrial machinery.	Sensor Streams	Learns spatio-temporal invariants to isolate rare failure modes natively.	High	Substantial parameter count requires dense data for stable optimization.
Transformer	Deep Learning	Uses multi-head self-attention to relate all elements in a sequence simultaneously.	Sequences	18 Captures highly complex long-range temporal correlations of normality.	Very High	Quadratic computational complexity relative to the input sequence length.
FTSGAN	Generative	Maps representations into fractional Fourier domain via adversarial training.	Spectral, Images	Regularizes high-frequency transitions to prevent collapse onto normal manifold.	High	Requires careful spectral domain calibration and complex training schedules.

of approximately 578:1, where the positive anomaly class accounts for only 0.172% of the dataset [63]. This extreme ratio disparity makes supervised model training highly susceptible to majority-class bias. To address this, hybrid frameworks frequently combine oversampling techniques like SMOTE-Tomek with sequential deep learning models [2, 4, 48]. This approach leverages bidirectional long short-term memory (BiLSTM) layers to capture subtle temporal dependencies in transaction sequences, while utilizing gated recurrent unit (GRU) blocks to preserve computational efficiency under dense transaction flows [3]. Despite these developments, transaction distributions remain highly non-stationary due to rapid shifts in consumer spending habits and evolving fraud strategies, known as concept drift, which degrades model performance over time.

5.2 Industrial Fault Detection and AIoT Diagnostics

In industrial engineering, real-time fault detection is vital to ensure structural integrity, prevent catastrophic machine breakdown, and minimize costly unplanned operational downtime. Class imbalance is a natural and structural property in this domain, as industrial assets are engineered to operate reliably, meaning that anomalies or failures occur with extremely low frequency. To evaluate model capabilities in these environments, several datasets are widely used, though their structural imbalance profiles differ significantly depending on the measurement granularity.

The Skoltech Anomaly Benchmark (SKAB) provides multivariate time series collected from a controlled water-pump testbed [64]. In SKAB, the anomaly distribution exhibits a dual nature depending on the evaluation objective:

- **Operational Duration Level (Continuous Time-Steps):** Across sustained anomalous operating windows, anomalous time-steps constitute 34.9% of the test records, yielding an apparent imbalance ratio of approximately 1.9:1.
- **Fault Onset Change point Level (Discrete Events):** At the discrete initiation boundary where physical degradation first occurs, there are only 129 change points across 37,401 records (0.35%), producing a severe event-level imbalance ratio of 289:1 [64].

Sound-based inspection is benchmarked using the Malfunctioning Industrial Machine Investigation and Inspection (MIMII) dataset, which contains recordings of pumps, valves, fans, and slide rails under normal and abnormal operating conditions [65]. MIMII exhibits an imbalance ratio of approximately 5:1 to 10:1, with normal operational audio spanning up to 10,000 seconds while faulty machine sounds are capped at roughly 1,000 seconds [65]. Conversely, the Case Western Reserve University (CWRU) bearing dataset serves as a standard reference for vibration analysis, though its original laboratory composition contains an unrealistic surplus of seeded fault classes relative to healthy data [5]. Consequently, researchers must selectively sub-sample CWRU to simulate the severe imbalance ratios ($> 10 : 1$ to $100 : 1$) encountered in real-world factories.

Recent breakthroughs in industrial Artificial Intelligence of Things (AIoT) have successfully mitigated these operational skews. Nguyen-Da et al. [6] and Nguyen-Thanh and Cho [7] deployed integrated CNN-LSTM and CNN-BiLSTM networks to monitor industrial diesel generators, demonstrating that localized spatial filtering combined with recurrent temporal tracking natively isolates rare mechanical failures without requiring risky synthetic oversampling. Furthermore, Nguyen-Thanh [49] established that coupling metaheuristic parameter optimization (IPSO) with hybrid deep networks enables early fault diagnosis in firefighting pumps under severe operational noise and sparse anomaly labels.

5.3 Healthcare and Medical Telemetry

In the medical domain, automated anomaly detection is critical for diagnostic support, patient triage, and continuous monitoring, where delayed or missed detections carry fatal consequences. Class imbalance in clinical settings is exceptionally pronounced because severe pathological conditions have a low prevalence compared to healthy cohorts. For example, in real-time intensive care unit monitoring, early sepsis prediction represents a major challenge due to high clinical skew. In the PhysioNet/Computing in Cardiology Challenge 2019 dataset, which contains clinical telemetry from 40,336 intensive care patients [66], class imbalance manifests at two distinct operational tiers:

- **Patient-Level Prevalence:** Across the total cohort of 40,336 patients, only 2,932 patients develop sepsis (7.27%), yielding a patient-level imbalance ratio of roughly 13.7:1 (or 12.8:1 depending on administrative cohort exclusion).
- **Hourly-Label Skew:** Across the continuous longitudinal telemetry comprising over 1,552,210 hourly records, only approximately 27,900 hourly observations (1.80%) are annotated as active septic onset windows, resulting in an extreme temporal imbalance ratio of 54.6:1 ($\approx 54 : 1$ or $1 : 53$) [66].

Similarly, rare disease diagnosis datasets often exhibit extreme imbalance ratios exceeding 100:1 or 500:1 due to the absolute scarcity of clinical samples for idiosyncratic conditions [24]. For physiological signal analysis, the MIT-BIH Arrhythmia database is the standard benchmark for electrocardiogram heartbeat anomaly detection, where normal sinus rhythm beats represent over 75% of the dataset, while rare arrhythmia classes, such as ventricular premature beats (PVC) or supraventricular premature beats, account for less than 1% of the samples, resulting in individual imbalance ratios exceeding 100:1 [67]. To address these severe imbalances, cost-sensitive convolutional neural networks, variational autoencoders, and generative adversarial networks dominate the medical literature, allowing researchers to penalize missed diagnoses heavily or synthesize realistic physiological signals [47].

5.4 Cybersecurity and Network Intrusion

In cybersecurity, network intrusion detection systems are vital to safeguard digital infrastructure from malicious exploits, denial of service attacks, and unauthorized

system access. In this domain, class imbalance is a persistent and evolving challenge because malicious network flows represent a minuscule fraction of overall network traffic, which is overwhelmingly dominated by benign packets. The legacy KDD Cup 99 dataset and its refined successor, the NSL-KDD dataset, serve as classic benchmarks, where multi-class distributions are severely skewed [68]. In NSL-KDD (comprising 148,517 total records across training and testing partitions), the imbalance must be understood across three analytical levels:

1. **Binary Attack-to-Normal Level:** Across the full dataset, normal traffic (77,054 records) and aggregate attack traffic (71,463 records) are broadly balanced, exhibiting an overall ratio of approximately 1.08:1 to 1.15:1.
2. **Categorical Attack Tier:** In multi-class evaluations, high-frequency attacks such as Neptune (DoS) encompass tens of thousands of flows, whereas intermediate minority categories such as Remote-to-Local (R2L, with 995 training flows) exhibit local imbalance ratios around 46:1 relative to the dominant Denial of Service clusters (45,927 flows) and 67:1 relative to normal traffic.
3. **Extreme Rare Attack Tier:** Highly dangerous exploits such as User-to-Root (U2R) contain only 52 samples in the entire 125,973-sample training set (0.041%), producing an extreme class-specific imbalance ratio of 1,295:1 relative to the 67,343 normal training instances [68].

This hierarchical skew is mirrored in modern benchmarks such as CICIDS2017, which exhibits an aggregate distribution of 80.3% benign flows and 19.7% malicious flows [9]. However, fine-grained exploits such as Infiltration or Heartbleed contain fewer than 50 flows amidst millions of benign records, leading to local imbalance ratios exceeding 10,000:1 [9].

5.5 Video Surveillance Anomaly Detection

Video surveillance anomaly detection is critical for maintaining public safety, optimizing emergency response, and securing critical infrastructure by automatically identifying abnormal events such as fights, vehicle intrusions, or thefts in camera feeds. Class imbalance is a defining characteristic of this domain, as surveillance feeds capture hours of completely benign, repetitive background activity, while anomalous events occur sporadically and last only a few seconds. Standard benchmarks are trained in an unsupervised, one-class manner using only normal video sequences. The UCSD Pedestrian dataset (Ped1 and Ped2) captures pedestrian walkways where abnormal events, such as vehicles or cyclists, are rare [69]. The CUHK Avenue dataset records campus pathways with anomalies like running, loitering, or throwing objects [70]. The ShanghaiTech Campus dataset represents the largest and most complex multi-scene benchmark, providing over 270,000 normal frames for training, while its test set contains only 130 anomalous events across 15,328 anomalous frames out of 317,371 test frames (4.8% anomaly fraction), yielding an operational test imbalance ratio of approximately 20:1 to 21:1 [10]. To capture these localized anomalies, spatiotemporal

autoencoders, 3D convolutional neural networks, and self-attention Transformers dominate the literature, mapping spatial textures alongside temporal motion trajectories [10, 52].

6 Benchmark Datasets and Evaluation Protocols

6.1 Common Benchmark Datasets

Evaluating anomaly detection frameworks under severe class imbalance requires standardized and diverse benchmark datasets. Table 4 organizes these key benchmark datasets to illustrate their scale, domain, and specific operational skew.

Table 4 Summary of standard benchmark datasets for imbalanced anomaly detection.

Dataset	Application Domain	Total Instances	Anomaly Fraction	Imbalance Ratio	Evaluation Formulation
Kaggle Fraud	Financial Transactions	284,807 transactions	0.172%	578:1	Semi-supervised / Binary
SKAB	Industrial Telemetry	37,401 time-steps	34.9% (sustained)	1.9:1 (duration)	Continuous observation evaluation
	(Fault Change-points)	129 onset events	0.35% (initiation)	289:1 (onset)	Change-point event detection
MIMII	Industrial Sound	10,000s audio files	10.0%–16.6%	5:1 to 10:1	One-class / Acoustic inspection
PhysioNet 2019	Clinical Patient Telemetry	40,336 patients	7.27% (patient cohort)	13.7:1 (patient)	Patient-level sepsis diagnosis
	Longitudinal Telemetry	1,552,210 hours	1.80% (hourly labels)	54:1 (hourly)	Streaming hourly-window triage
MIT-BIH	Cardiac Telemetry	109,446 beats	< 1.0% (rare classes)	> 100 : 1	Supervised Multi-class heartbeat
NSL-KDD	Network Traffic (Binary)	148,517 flows	46.5% (total attacks)	1.15:1 (binary)	Attack vs. normal baseline
	Intra-Attack Tail (U2R)	125,973 train flows	0.041% (52 samples)	1,295:1 (U2R class)	Rare exploit isolation
CICIDS2017	Network Intrusion	2,830,743 flows	< 0.01% (Infiltration)	> 10,000 : 1	Multi-class / Streaming packets
ShanghaiTech	Video Surveillance	270,000 train / 317,000 test	4.8% (test frames)	20:1	Unsupervised / Spatio-temporal

6.2 Evaluation Pitfalls in Imbalanced Anomaly Detection

The integrity of imbalanced anomaly detection research is compromised by critical, recurring evaluation pitfalls:

6.2.1 Pre-Partition Resampling Leakage

The most rampant and mathematically fatal error is data leakage occurring when resampling techniques are executed prior to dataset partitioning. If synthetic oversampling techniques like SMOTE or ADASYN are applied to the entire dataset before performing a train-test split, the coordinates of the synthetic instances in the training set are interpolated using features from the future test set:

$$x_{\text{synthetic}} = x_{\text{train}} + \lambda(x_{\text{test}} - x_{\text{train}}) \quad (13)$$

This leaks the testing distribution directly into the training phase, essentially giving the model a cheat sheet. Consequently, during evaluation, the model classifies clone-like test instances with deceptive near-perfection, producing inflated F1-scores (> 0.98) that collapse completely when deployed in live production environments. Preprocessing steps, such as min-max normalization and missing value imputation, must similarly be calculated strictly on the training partition and subsequently applied as a fixed transform to the test partition to prevent feature leakage.

6.2.2 The Accuracy Paradox and Single-Metric Illusion

As established in Section 3.3, overall classification accuracy is completely uninformative in skewed domains. Models optimizing accuracy converge to the majority-class trivial attractor. Furthermore, relying solely on F1-score can be misleading if the classification threshold is tuned post-hoc on the test set rather than a dedicated validation set. Threshold tuning on test data introduces optimistic bias, misrepresenting operational precision-recall trade-offs.

6.2.3 Temporal Contamination and Shuffling in Time-Series

Finally, time-series telemetry (such as industrial vibration, medical telemetry, and financial transaction streams) suffers from temporal train-test contamination when split randomly rather than chronologically. In transactional streams, a random split allows future transaction profiles to leak into past predictions, violating temporal causality and ignoring natural concept drift. To prevent these systematic errors, researchers must enforce chronological data splits and integrate all data scaling and resampling steps strictly within localized cross-validation loops.

6.3 Quantitative Performance Synthesis

To provide a consolidated benchmark reference across diverse methodological paradigms, Table 5 synthesizes quantitative performance metrics reported in state-of-the-art literature across all five discussed application domains.

Table 5 Quantitative performance synthesis across benchmark datasets and paradigms.

Method	Family	Dataset (Domain)	Imb. Ratio	Prec.	Rec.	F1	AUC
XGBoost (Cost-Weighted)	Classical ML	Kaggle Fraud (Financial)	578:1	0.865	0.792	0.827	0.974
Autoencoder	Deep One-Class	Kaggle Fraud (Financial)	578:1	0.810	0.860	0.830	0.920
BiLSTM + GRU (SMOTE-Tomek)	Hybrid Sequential	Kaggle Fraud (Financial)	578:1	0.920	0.950	0.930	0.990
1D CNN	Deep Learning	CWRU Bearing (Industrial)	> 10 : 1	0.950	0.930	0.940	0.960
Cost-Sensitive SVM	Algorithm-Level	CWRU Bearing (Industrial)	> 10 : 1	0.910	0.890	0.900	0.930
CNN-BiLSTM (IPSO-tuned)	AIoT Hybrid	Industrial AIoT (Industrial)	> 25 : 1	0.942	0.938	0.940	0.968
Random Forest (Balanced)	Classical ML	NSL-KDD (Security)	46:1	0.890	0.810	0.850	0.900
Focal-Loss Deep Net	Algorithm-Level	CICIDS2017 (Security)	> 1,000 : 1	0.873	0.841	0.857	0.946
Cost-Sensitive CNN	Algorithm-Level	MIT-BIH (Healthcare)	> 100 : 1	0.884	0.871	0.877	0.952
TCN-GRU Architecture	Sequential Deep	PhysioNet 2019 (Healthcare)	54:1	0.745	0.789	0.766	0.885
3D Conv-Autoencoder	Deep One-Class	ShanghaiTech (Surveillance)	20:1	–	–	–	0.728
Spatio-Temporal Transformer	Self-Attention	ShanghaiTech (Surveillance)	20:1	–	–	–	0.784

7 Open Challenges and Future Directions

7.1 Real-Time Streaming Anomaly Detection under Concept Drift

Adapting anomaly detection models to real-time streaming data environments remains a persistent research challenge due to strict memory and latency constraints. In streaming time-series pipelines, such as IoT sensor arrays or high-frequency stock trading, traditional batch-based data resampling techniques like SMOTE are completely unusable because they require offline access to the complete feature space. This limitation is compounded by concept drift, where the underlying statistical distribution of both normal and anomalous data changes over time. Existing drift detection algorithms, including adaptive windowing, are mathematically designed under the assumption of balanced class distributions. Under extreme class imbalance, these detectors are heavily biased toward the dominant class, frequently misinterpreting minor distribution

changes in normal data as anomalies, while failing to identify actual semantic shifts in rare target events [71]. Developing online learning algorithms that can dynamically adjust decision boundaries in highly skewed, drifting data streams without undergoing catastrophic forgetting is a critical future path.

7.2 Federated Learning with Non-IID Skewed Local Distributions

Federated learning offers a decentralized training paradigm that preserves data privacy, yet it faces unique optimization challenges when deployed under severe class skews. In practical federated setups, such as collaborative medical diagnostics across multiple hospitals, class imbalance is highly heterogeneous across edge clients. Some local clients may possess relatively balanced local datasets, while other clients suffer from extreme 1,000:1 imbalances or even exhibit vacant classes where specific anomalous modes are completely absent. Standard federated aggregation protocols, such as Federated Averaging, undergo severe gradient divergence because local models optimize along highly biased, client-specific loss surfaces. This gradient divergence degrades the performance of the aggregated global model and slows overall convergence. Furthermore, classical data-level resampling methods cannot be deployed globally without violating strict privacy constraints, leaving local clients unable to align their decision boundaries collaboratively. Mitigating this local imbalance disparity without sharing raw edge data represents an important, unresolved open challenge.

7.3 Explainability and Interpretability Breakdown under Minority Sparsity

In safety-critical and highly regulated applications, anomaly detection models must provide transparent, auditable decision paths to be trusted by human operators. However, standard post-hoc explainability techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), exhibit severe instability and complete interpretability breakdown when applied to highly imbalanced datasets. Because these model-agnostic explanation tools rely on localized feature perturbations to calculate local feature attributions, the extreme sparsity of the minority class causes these generated perturbations to fall almost entirely within the majority class manifold. This sampling bias produces inconsistent and deceptive feature importance scores that do not reflect the actual decision boundaries of the classifier for rare events. Consequently, security analysts and clinicians are often presented with unreliable explanations for critical positive detections, eroding trust. Bridging the gap between predictive performance on skewed distributions and robust explanation stability remains a key future direction.

7.4 Active Human-in-the-Loop Anomaly Discovery

Unsupervised and semi-supervised anomaly detection algorithms are prone to elevated false positive rates, which quickly causes operator alert fatigue in industrial and cybersecurity monitoring centers. Human-in-the-loop systems aim to mitigate this

operational overhead by utilizing active learning query strategies to systematically refine decision boundaries based on human expert feedback [72]. However, designing active query strategies under extreme class imbalance is exceptionally difficult. Traditional uncertainty sampling heuristics tend to query instances that lie closest to the current decision boundary, which in highly imbalanced settings are overwhelmingly dominated by ambiguous majority class instances rather than actual anomalies. This misdirection wastes limited human verification effort on benign noise. Designing specialized active query frameworks that can discover diverse anomaly modes while systematically minimizing human verification overhead remains an open, unresolved problem.

7.5 Cross-Domain Generalization and Representation Decoupling

Catastrophic model failure occurs when anomaly detection frameworks are transferred to a target domain with a similar imbalance ratio but different feature distributions or schemas. In cross-domain graph or tabular anomaly detection, models struggle because the source domain decision boundary is tightly overfit to the highly specific local majority-class structure. Standard unsupervised domain adaptation methods rely on class-balanced source distributions to align source and target representations. When applied to highly imbalanced datasets, the adversarial domain alignment process frequently aligns the dense source majority class with the sparse target minority class, destroying the model capacity to isolate target anomalies. Developing robust, cross-domain representation learning techniques, such as graph transfer and graph decouple architectures [58], that can align domain-invariant features while preserving the distinct, rare signatures of anomalies represents a significant, unaddressed research gap.

7.6 Large Language Model (LLM) Integration in Extreme Anomaly Reasoning

Finally, utilizing Large Language Models (LLMs) for zero-shot tabular or text log anomaly detection represents a nascent research direction. Specifically, LLMs face significant tokenization limitations when handling continuous-valued tabular and multi-dimensional time-series data. Serializing dense float arrays into raw text representations destroys local numerical precision, erodes topological feature proximity, and drastically inflates context window consumption. Moreover, the autoregressive generation latency of LLMs is completely incompatible with microsecond-level real-time networks required for high-frequency financial fraud detection or critical cyber-physical industrial diagnostics. This extreme compute overhead makes LLMs highly impractical compared to lightweight classical baselines like XGBoost or shallow deep autoencoders. Nevertheless, promising hybrid research directions are emerging. Rather than utilizing LLMs for raw sample-level detection, recent frameworks deploy them as downstream explainers that ingest pre-flagged anomalous telemetry to generate natural language root-cause descriptions. Alternatively, prompt-based zero-shot anomaly scoring has shown utility in text-rich log environments, such as system security logs or Kubernetes

events, where semantic reasoning over unstructured metadata is crucial and real-time execution speeds are not strictly bottlenecked by line-rate constraints.

8 Conclusion

This survey has synthesized the complex research landscape at the intersection of anomaly detection and extreme class imbalance. By systematically organizing data-level, algorithm-level, classical machine learning, deep learning, generative, and self-supervised paradigms, we demonstrated that resolving severe data skews requires moving beyond naive global loss optimization. Our cross-domain evaluation across financial, industrial, clinical, security, and surveillance sectors highlighted that model efficacy is intrinsically bound to operational constraints, such as noise tolerance, latency, and label availability.

Crucially, our audit exposed that methodological rigor is frequently undermined by flawed evaluation practices. The pervasive occurrence of data leakage during pre-partition resampling, the persistence of the accuracy paradox, and the temporal contamination of streaming data splits produce dangerously optimistic benchmark scores that fail upon real-world deployment. Researchers and practitioners must rigorously enforce chronological partitioning, within-fold preprocessing, and threshold-free minority evaluation metrics (such as AUC-PR and MCC).

Looking forward, the community must focus on three core research frontiers: (1) low-latency, online learning architectures capable of detecting concept drift in highly imbalanced continuous streams; (2) privacy-preserving federated aggregation mechanisms that tolerate extreme local client skew; and (3) explanation frameworks that remain stable despite extreme minority sparsity. Addressing these foundational challenges will enable the development of truly dependable, deployable anomaly detection systems in skewed real-world data environments.

Declarations.

- **Funding:** No external funding was received for conducting this study.
- **Conflict of Interest:** The author declares no competing financial or non-financial interests directly or indirectly related to the work submitted for publication.
- **Availability of Data and Materials:** All benchmark datasets synthesized in this survey (Kaggle Credit Card Fraud, SKAB, MIMII, PhysioNet 2019, MIT-BIH, NSL-KDD, CICIDS2017, and ShanghaiTech) are publicly available through their referenced academic repositories.
- **Author Contributions:** Muhammad Jabir Khan was solely responsible for conceptualization, methodology, literature curation, formal analysis, manuscript drafting, visual illustrations, and revisions.

References

- [1] Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN Computer Science **2**(3), 160 (2021) <https://doi.org/10.1007/>

- [2] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* **6**(1), 20–29 (2004) <https://doi.org/10.1145/1007730.1007735>
- [3] Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734 (2014)
- [4] Khan, M.J.: Addressing class imbalance in credit card fraud detection: A hybrid deep learning approach. *International Journal of Innovations in Science and Technology* **7**(1), 603–622 (2025)
- [5] Ren, Z., Lin, T., Feng, K., Zhu, Y., Liu, Z., Yan, K.: A systematic review on imbalanced learning methods in intelligent fault diagnosis. *IEEE Transactions on Instrumentation and Measurement* **72**, 1–35 (2023) <https://doi.org/10.1109/TIM.2023.3285973>
- [6] Nguyen-Da, T., Nguyen-Thanh, P., Cho, M.-Y.: Real-time AIoT anomaly detection for industrial diesel generator based an efficient deep learning CNN-LSTM in industry 4.0. *Internet of Things* **27**, 101280 (2024) <https://doi.org/10.1016/j.iot.2024.101280>
- [7] Nguyen-Thanh, P., Cho, M.-Y.: Advanced AIoT for failure classification of industrial diesel generators based hybrid deep learning CNN-BiLSTM algorithm. *Advanced Engineering Informatics* **62**, 102644 (2024) <https://doi.org/10.1016/j.aei.2024.102644>
- [8] Fernando, T., Gammulle, H., Sridharan, S., Fookes, C.: Deep learning for medical anomaly detection – a survey. *ACM Computing Surveys* **54**(7), 1–37 (2021) <https://doi.org/10.1145/3464423>
- [9] Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: Toward generating a new dataset for intrusion detection to detect novel attacks. In: *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (SECURITY)*, pp. 108–116 (2018)
- [10] Luo, W., Liu, W., Gao, S.: A revisit of video anomaly detection by deep reconstruction joint with representation learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3111–3120 (2017)
- [11] He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284 (2009) <https://doi.org/10.1109/TKDE.2009.126>

- [12] Krawczyk, B.: Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence* **5**(4), 221–232 (2016) <https://doi.org/10.1007/s13748-016-0094-0>
- [13] Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., Japkowicz, N.: The class imbalance problem in deep learning. *Machine Learning* **113**(7), 4845–4901 (2024) <https://doi.org/10.1007/s10994-024-06541-7>
- [14] Chen, W., Yang, K., Yu, Z., Shi, Y., Chen, C.L.P.: A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review* **57**(6), 137 (2024) <https://doi.org/10.1007/s10462-024-10758-0>
- [15] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. *ACM Computing Surveys* **41**(3), 1–58 (2009) <https://doi.org/10.1145/1541880.1541882>
- [16] Pang, G., Shen, C., Cao, L., Hengel, A.: Deep learning for anomaly detection: A review. *ACM Computing Surveys* **54**(2), 1–38 (2021) <https://doi.org/10.1145/3439950>
- [17] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Moher, D.: The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **372**, 71 (2021) <https://doi.org/10.1136/bmj.n71>
- [18] Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-2**(3), 408–421 (1972)
- [19] Tomek, I.: Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-6**(11), 769–772 (1976)
- [20] Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. *Neural Computation* **13**(7), 1443–1471 (2001) <https://doi.org/10.1162/089976601750264965>
- [21] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002) <https://doi.org/10.1613/jair.953>
- [22] Powers, D.M.W.: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2**(1), 37–63 (2011)
- [23] Fawcett, T.: An introduction to ROC analysis. *Pattern Recognition Letters* **27**(8), 861–874 (2006) <https://doi.org/10.1016/j.patrec.2005.10.010>

- [24] Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**(1), 6 (2020) <https://doi.org/10.1186/s12864-019-6413-7>
- [25] Kim, B., Kim, J.: Adjusting decision boundary for class imbalanced learning. *IEEE Access* **8**, 81674–81685 (2020) <https://doi.org/10.1109/ACCESS.2020.2991196>
- [26] Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: *International Conference on Intelligent Computing (ICIC)*. LNCS, vol. 3644, pp. 878–887 (2005)
- [27] Nguyen, H.M., Cooper, E.W., Kamei, K.: Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms* **3**(1), 4–21 (2011)
- [28] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1322–1328 (2008)
- [29] Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery* **28**(1), 92–122 (2014) <https://doi.org/10.1007/s10618-012-0295-5>
- [30] Zhang, J., Mani, I.: kNN approach to unbalanced data distributions: A case study involving information extraction. In: *Proceedings of the ICML Workshop on Learning from Imbalanced Datasets*, pp. 1–7 (2003)
- [31] Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A Python toolbox to tackle the class imbalance problem in machine learning. *Journal of Machine Learning Research* **18**(17), 1–5 (2017)
- [32] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 7335–7345 (2019)
- [33] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 2672–2680 (2014)
- [34] Elkan, C.: The foundations of cost-sensitive learning. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–978 (2001)
- [35] Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the sensitivity of support vector machines. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 55–60 (1999)

- [36] Domingos, P.: Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155–164 (1999)
- [37] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
- [38] Tax, D.M.J., Duin, R.P.W.: Support vector data description. *Machine Learning* **54**(1), 45–66 (2004) <https://doi.org/10.1023/B:MACH.0000008084.60811.49>
- [39] Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data* **6**(1), 1–39 (2012) <https://doi.org/10.1145/2133360.2133363>
- [40] Liu, X.-Y., Wu, J., Zhou, Z.-H.: Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(2), 539–550 (2009) <https://doi.org/10.1109/TSMCB.2008.2007853>
- [41] Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* **40**(1), 185–197 (2010) <https://doi.org/10.1109/TSMCA.2009.2029569>
- [42] King, G., Zeng, L.: Logistic regression in rare events data. *Political Analysis* **9**(2), 137–163 (2001) <https://doi.org/10.1093/oxfordjournals.pan.a004868>
- [43] Batuwita, R., Palade, V.: Class imbalance learning methods for support vector machines. In: *Imbalanced Learning: Foundations, Algorithms, and Applications*, pp. 83–99. Wiley-IEEE Press, ??? (2013)
- [44] Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (2001) <https://doi.org/10.1023/A:1010933404324>
- [45] O’Brien, R., Ishwaran, H.: A random forests quantile classifier for class imbalanced data. *Pattern Recognition* **90**, 232–249 (2019) <https://doi.org/10.1016/j.patcog.2019.01.036>
- [46] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
- [47] Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: *International Conference on Learning Representations (ICLR)*, pp. 1–14 (2014)
- [48] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation* **9**(8), 1735–1780 (1997) <https://doi.org/10.1162/neco.1997.9.8.1735>

- [49] Nguyen-Thanh, P.: Firefighting pump anomaly analysis in AIoT maintenance framework based on IPSO and BiLSTM-CNN deep learning approaches. *Software: Practice and Experience* **56**(5), 1–22 (2026) <https://doi.org/10.1002/spe.3289>
- [50] Nguyen, D.-T., Nguyen, T.-P., Cho, M.-Y.: Online prognostic failure AIoT system for industrial generators maintenance service based two-stage deep learning algorithm. *Control Engineering Practice* **157**, 106263 (2025) <https://doi.org/10.1016/j.conengprac.2024.106263>
- [51] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015) <https://doi.org/10.1038/nature14539>
- [52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008 (2017)
- [53] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: *International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 146–157 (2017)
- [54] Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis* **54**, 30–44 (2019) <https://doi.org/10.1016/j.media.2019.01.010>
- [55] Wang, D., Gao, L., Qu, Y., Sun, X., Liao, W.: FTSGAN: Frequency-to-spectrum mapping GAN for semisupervised hyperspectral anomaly detection. *CAAI Transactions on Intelligence Technology* **8**(4), 1258–1273 (2023) <https://doi.org/10.1049/cit2.12211>
- [56] Wang, D., Zhuang, L., Gao, L., Sun, X., Huang, M., Plaza, A.: BockNet: Blind-block reconstruction network with a guard window for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 1–16 (2023) <https://doi.org/10.1109/TGRS.2023.3289124>
- [57] Wang, D., Zhuang, L., Gao, L., Sun, X., Zhao, X.: Global feature-injected blind-spot network for hyperspectral anomaly detection. *IEEE Geoscience and Remote Sensing Letters* **21**, 1–5 (2024) <https://doi.org/10.1109/LGRS.2024.3361895>
- [58] Huang, C., Shi, X., Gao, C., Hu, Q., Huang, X., Huang, Q., Anaissi, A.: Cross-domain graph anomaly detection via graph transfer and graph decouple. *CAAI Transactions on Intelligence Technology* **10**(4), 1089–1103 (2025) <https://doi.org/10.1049/cit2.12351>
- [59] Mirza, M., Osindero, S.: Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014)

- [60] Tack, J., Yu, S., Jeong, J., Kim, M., Shin, J.: CSI: Novelty detection with contrastive learning on distributionally shifted instances. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 11839–11852 (2020)
- [61] Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 9758–9769 (2018)
- [62] Li, C.-L., Sohn, K., Yoon, J., Pfister, T.: CutPaste: Self-supervised learning for anomaly detection and localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9664–9674 (2021)
- [63] Pozzolo, A.D., Caelen, O., Johnson, R.A., Bontempi, G.: Calibrating probability with undersampling for unbalanced classification. In: *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 159–166 (2015)
- [64] Katser, I.D., Kozitsin, V.O.: Skoltech Anomaly Benchmark (SKAB). Kaggle Dataset (2020). <https://doi.org/10.34740/KAGGLE/DSV/1693952>
- [65] Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., Kawaguchi, Y.: MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pp. 209–213 (2019)
- [66] Reyna, M.A., Josef, C., Jeter, R., Nemati, S., Westover, M.B., Clifford, G.D., Malhotra, A.: Early prediction of sepsis from clinical data: The PhysioNet/computing in cardiology challenge 2019. In: *Computing in Cardiology (CinC)*, vol. 46, pp. 1–4 (2019)
- [67] Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine* **20**(3), 45–50 (2001)
- [68] Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: *IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pp. 1–6 (2009)
- [69] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1975–1981 (2010)
- [70] Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in Matlab. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2720–2727 (2013)

- [71] Pezzicoli, F.S., Baity-Jesi, M., the Exactly Solvable Models Group: Class imbalance in anomaly detection: Learning from an exactly solvable model. In: Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, vol. 258, pp. 1261–1269 (2025)
- [72] Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1), 1–114 (2012)