

A REVIEW OF COMPARATIVE STUDY OF COST-SENSITIVE LEARNING AND DATA-LEVEL BALANCING STRATEGIES FOR HIGHLY SKEWED SPAM EMAIL DATASETS

Jyoti Yadav¹, Mrs. Arifa Khan²

¹Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

²Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

Abstract - Class imbalance is a pervasive challenge in spam email classification, where legitimate messages overwhelmingly outnumber spam instances, leading to biased learning and degraded classifier performance. This review paper critically examines two major strategies for handling highly skewed spam email datasets: cost-sensitive learning and data-level balancing techniques. Cost-sensitive learning embeds misclassification costs directly into the learning process to penalize errors on minority classes, whereas data-level methods adjust the dataset distribution using sampling strategies such as oversampling, undersampling, and synthetic sample generation. Through a systematic synthesis of contemporary research, this review highlights the theoretical foundations, algorithmic implementations, and practical impacts of these approaches on popular spam datasets such as SpamBase, Enron, and LingSpam. Comparative evaluations reveal that while data-level methods like SMOTE variants and hybrid sampling can improve class representation, they may introduce noise or overfitting. Conversely, cost-sensitive frameworks offer flexible decision boundaries but often require careful cost matrix tuning. Direct comparative studies suggest that hybrid combinations of both strategies yield more robust performance across diverse imbalance ratios. The review also identifies key limitations in existing literature—including a lack of standardized benchmarking, limited examination of evolving spam patterns, and underexplored deep learning adaptations. Finally, future research directions are outlined to bridge current gaps and enhance spam detection in increasingly dynamic environments.

Key Words: Imbalanced Learning, Cost-Sensitive Classification, Data-Level Balancing, Spam Email Detection, Synthetic Oversampling

1. INTRODUCTION

1.1 Background

1.1.1 Importance of Email Spam Detection

Email spam detection remains a critical area of study in machine learning and cybersecurity due to the massive volume of unsolicited emails circulated daily and the potential risks these messages pose to users and network

infrastructures. Unwanted or malicious emails contribute not only to user frustration and productivity loss, but also to the propagation of phishing attempts, malware distribution, and data theft campaigns, which can compromise security and privacy across digital ecosystems (Fernández-López et al., 2018; Akeel et al., 2025). Traditional rule-based filters, which once dominated early spam filtering systems, often fail to cope with evolving content patterns and obfuscation techniques employed by spammers, necessitating the use of more advanced machine learning and NLP-driven classification methods (Akeel et al., 2025). Machine learning-based approaches improve adaptability by learning discriminative patterns from historical datasets, making them resilient to changing spam tactics and text variations (Asliyukse, Tonkal & Kocaoglu, 2025).

1.1.2 Challenges Posed by Highly Skewed Datasets (Class Imbalance)

One of the core challenges in spam classification is the class imbalance problem, where legitimate (non-spam) emails significantly outnumber spam instances. This imbalance leads standard learning algorithms to bias predictions towards the majority class, resulting in poor detection of the minority (spam) class and inflated performance metrics such as overall accuracy that mask poor recall on the minority category (Fernández-López et al., 2018). Such skewed distributions reduce the effective learning of decision boundaries for rare classes, while increasing false negatives — a critical issue in real-world deployments where missed spam can have serious consequences. To mitigate these issues, researchers propose both data-level balancing (resampling) and cost-sensitive learning strategies that adjust either the data distribution or the learning process to improve minority class detection without discarding valuable information.

1.2 Motivation for Review

1.2.1 Comparing Cost-Sensitive Learning and Data-Level Balancing Is Important

In the literature on imbalanced classification, two broad paradigms have emerged to address skewness: data-level techniques alter the dataset itself (e.g., oversampling,

undersampling, synthetic sample generation), whereas cost-sensitive learning embeds misclassification costs into algorithms to penalize errors on the minority class more heavily. Both strategies aim to reduce the bias introduced by imbalanced data, but they differ fundamentally in approach and implications for classifier performance (Feng, Zhou & Tong, 2020; Fernández-López et al., 2018). Data balancing has been shown to improve representation of rare classes by augmenting or undersampling data, but may introduce noise or information loss if not carefully tuned. Cost-sensitive frameworks focus on adjusting the decision criterion rather than altering the dataset, potentially preserving valuable information while explicitly incorporating error costs. However, the effectiveness of these methods varies with dataset characteristics, model choice, and the imbalance ratio, making comparative evaluation essential for understanding their relative strengths and trade-offs.

1.2.2 Limitations of Existing Surveys

Although there are surveys on class imbalance and general strategies for handling skewed datasets, many existing reviews lack focused comparative analysis specifically tailored to spam email classification datasets. Prior reviews often consider broad applications such as fraud detection or medical diagnosis, where imbalance characteristics and cost implications differ significantly from text-based spam problems. Moreover, few surveys integrate recent advances in synthetic sample generation or hybrid models that combine both cost-sensitive and balancing techniques, leaving a gap in comprehensive assessment of these strategies in the context of evolving spam tactics and textual features. This review aims to fill that gap by synthesizing current research and providing detailed comparisons informed by recent empirical outcomes.

1.3 Scope and Objectives

This paper systematically examines the literature on cost-sensitive and data-level imbalance handling strategies as applied to spam email datasets. It surveys key methods, including oversampling, undersampling, synthetic data techniques, and cost-sensitive algorithm adaptations, drawing insights from varied experimental evaluations and algorithm implementations. By comparing performance trends and observed trade-offs across different studies, the review highlights where each strategy shows promise or limitations, and synthesizes conclusions that can guide future research and practical applications.

2. PROBLEM DEFINITION

2.1 Spam Email Classification

2.1.1 Characteristics and Features of Spam vs. Legitimate Emails

Spam email classification is a text-based binary classification task that differentiates unsolicited or malicious emails (spam) from legitimate emails (ham). Spam emails often exhibit distinctive linguistic and structural patterns, including the use of promotional keywords, deceptive URLs, excessive capitalization, and unusual punctuation, whereas legitimate emails generally follow formal, contextually relevant structures (Asliyukse, Tonkal & Kocaoglu, 2025). Features commonly employed for classification include lexical attributes (word frequency, n-grams), semantic features derived from natural language processing (NLP), sender metadata, and email header information (Fernández-López et al., 2018). High-quality feature extraction is essential because spam techniques continuously evolve, with spammers using obfuscation methods such as text randomization, embedding URLs within images, and content variation to evade detection (Akeel et al., 2025). Accurate representation of these characteristics in model inputs is therefore critical for effective spam detection, particularly when dealing with large and heterogeneous email datasets.

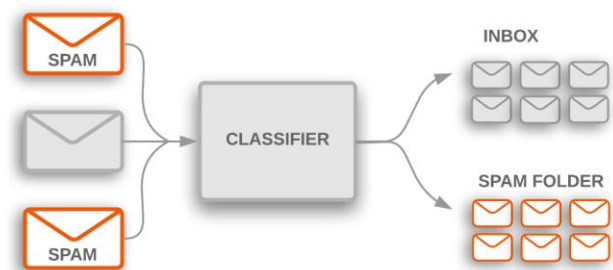


Figure-1: Spam Email Classification

2.2 Highly Skewed Datasets

2.2.1 Nature and Implications of Class Imbalance

In spam detection datasets, legitimate emails usually outnumber spam emails by a significant margin, leading to highly skewed or imbalanced class distributions. This imbalance can bias standard machine learning algorithms toward predicting the majority class, resulting in a disproportionate number of misclassified minority instances (Feng, Zhou & Tong, 2020). Imbalance also reduces the ability of classifiers to learn discriminative patterns for the minority class, which is particularly problematic in real-world scenarios where spam emails may carry phishing or malware risks.

2.2.2 Performance Metrics Affected by Imbalance

Traditional evaluation metrics like overall accuracy can be misleading under class imbalance, often suggesting high performance even when the classifier fails to correctly identify spam instances (Fernández-López et al., 2018). Metrics that account for imbalance, such as precision, recall, F1-score, area under the ROC curve (AUC), G-mean, and Matthews correlation coefficient (MCC), provide a more realistic assessment of classifier effectiveness for the minority class. These metrics help highlight the trade-offs between false positives (legitimate emails misclassified as spam) and false negatives (spam emails missed by the classifier), which are crucial in practical deployments.

2.3 Challenges Specific to Spam Datasets

2.3.1 High Dimensionality

Spam datasets often include a very large number of textual and metadata features, leading to high-dimensional data spaces. High dimensionality increases computational complexity, risks overfitting, and can dilute the effectiveness of traditional machine learning algorithms if irrelevant or noisy features are not appropriately filtered (Akeel et al., 2025).

2.3.2 Evolving Spamming Patterns

Spammers constantly adapt their methods to bypass detection systems, introducing evolving patterns that complicate classification. These may include dynamic subject lines, changing sender addresses, and content obfuscation strategies, which require classifiers to generalize effectively across unseen spam types and maintain robustness over time (Asliyukse, Tonkal & Kocaoglu, 2025).

2.3.3 Data Sparsity

Despite the large number of emails in real-world datasets, features relevant to spam detection—such as rare keywords, uncommon URLs, or embedded malware indicators—can be sparse. Sparse data reduces statistical strength for learning, particularly for minority class instances, and necessitates specialized methods like oversampling, synthetic data generation, or cost-sensitive learning to improve classifier performance (Feng, Zhou & Tong, 2020).

3. IMBALANCE IN MACHINE LEARNING

3.1 Class Imbalance: General Concept

3.1.1 Definitions

Class imbalance in machine learning refers to scenarios where the number of instances in one class (typically the majority) significantly exceeds those in another (minority), creating a skewed distribution (Fernández-López et al., 2018). This imbalance is prevalent in real-world domains

such as fraud detection, medical diagnosis, and spam email classification, where minority instances carry higher importance. In the context of spam detection, the majority class typically consists of legitimate emails, while spam emails represent the minority class, often comprising less than 10% of the dataset.

3.1.2 Impact on Classifier Performance

Imbalanced datasets pose a significant challenge to standard machine learning algorithms, which tend to favor the majority class. This can result in classifiers exhibiting high overall accuracy while performing poorly on the minority class, leading to false negatives that are particularly detrimental in spam detection, where missed spam emails may carry malware or phishing threats (Feng, Zhou & Tong, 2020). Furthermore, imbalance can complicate the learning of discriminative boundaries, reduce model generalization, and exacerbate overfitting on the minority class if improper balancing strategies are applied.

3.2 Evaluation Metrics for Imbalanced Datasets

3.2.1 Confusion Matrix

The confusion matrix is a foundational tool for evaluating classification performance under imbalance, as it provides a detailed breakdown of true positives, true negatives, false positives, and false negatives. This granular view enables researchers to identify biases toward the majority class and assess how well minority instances are captured.

3.2.2 Precision, Recall, and F1-Score

Metrics such as precision, recall, and F1-score are more informative than overall accuracy in imbalanced scenarios (Fernández-López et al., 2018). Precision indicates the proportion of correctly identified positive instances among all predicted positives, while recall measures the proportion of correctly identified positive instances relative to all actual positives. The F1-score, as the harmonic mean of precision and recall, balances both aspects and is particularly useful for evaluating minority-class detection.

3.2.3 AUC-ROC, G-Mean, MCC

Advanced metrics like the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), geometric mean (G-Mean), and Matthews Correlation Coefficient (MCC) offer more comprehensive performance assessment for imbalanced datasets. AUC-ROC evaluates classifier discrimination capability independent of threshold, G-Mean balances sensitivity and specificity, and MCC accounts for all four confusion matrix elements, making it robust against skewed distributions (Feng, Zhou & Tong, 2020).

3.2.4 Cost-Sensitive Metrics

In addition, cost-sensitive metrics assign different penalties to misclassifications of minority versus majority classes. These metrics are essential in contexts like spam detection, where the cost of missing a spam email (false negative) may outweigh incorrectly flagging a legitimate email (false positive) (Akeel et al., 2025). Incorporating these metrics helps guide algorithm design and evaluation.

3.3 Benchmark Spam Datasets with Imbalance

3.3.1 UCI SpamBase, Enron, LingSpam

Several benchmark datasets are widely used for research in imbalanced spam email classification. UCI SpamBase contains 4,601 emails with a 39.4% spam ratio, featuring word frequency, character frequency, and capital letter statistics. Enron Email Dataset includes millions of real-world emails with extensive metadata and an approximate 13% spam ratio, providing opportunities for large-scale evaluation. LingSpam Dataset comprises around 2,893 emails, with roughly 23% spam, emphasizing textual content and structural features.

3.3.2 Feature Characteristics and Imbalance Ratios

These datasets vary in terms of feature types and imbalance ratios, which directly affect algorithm performance and the choice of imbalance mitigation strategies. UCI SpamBase emphasizes token-level lexical features, Enron offers metadata and social network features, while LingSpam focuses on linguistic patterns. Understanding these characteristics is crucial for selecting appropriate resampling methods, cost-sensitive algorithms, or hybrid approaches.

4. DATA-LEVEL BALANCING STRATEGIES

Data-level balancing strategies are widely used techniques to address class imbalance by modifying the training dataset rather than changing the learning algorithm itself (Fernández-López et al., 2018). These methods aim to provide classifiers with more representative examples of the minority class, thereby improving prediction performance. They are particularly relevant in spam email classification, where minority spam messages are often underrepresented and difficult to detect accurately.

4.1 Random Oversampling

4.1.1 Basic Approach

Random oversampling increases the number of minority class instances by randomly duplicating existing samples in the training dataset (Feng, Zhou & Tong, 2020). This ensures that the minority class is more equally represented during

model training, which can improve the ability of classifiers to recognize rare spam emails.

4.1.2 Drawbacks (Overfitting)

While oversampling is simple and effective, it can lead to overfitting, as the classifier may memorize duplicated minority samples instead of learning generalized patterns. This is especially problematic in high-dimensional spam datasets, where duplicated emails may not contribute new information but still bias the model (Akeel et al., 2025).

4.2 Random Undersampling

4.2.1 Trade-offs (Loss of Information)

Random undersampling reduces the number of majority class instances to balance the dataset. Although it can mitigate bias toward the majority class and reduce training time, it may result in loss of critical information, as legitimate emails that carry diverse patterns are discarded (Fernández-López et al., 2018). For spam datasets with limited minority instances, this can negatively impact the model's generalization performance.

4.3 Synthetic Oversampling

4.3.1 SMOTE and Its Variants

Synthetic Minority Oversampling Technique (SMOTE) generates new minority class samples by interpolating between existing instances in the feature space (Feng, Zhou & Tong, 2020). Variants such as Borderline-SMOTE focus on generating synthetic samples near class boundaries, while ADASYN (Adaptive Synthetic Sampling) generates more samples in regions where the minority class is harder to learn.

4.3.2 How Synthetic Generation Works

Synthetic generation works by selecting a minority class instance and creating new examples along the line segments connecting it to its nearest minority neighbors. This approach increases diversity among minority samples, reducing overfitting compared to random oversampling.

4.3.3 Application to Spam Datasets

In spam email datasets, SMOTE and its variants have been applied to augment rare spam messages, improving classifier recall and F1-score without significantly affecting precision (Akeel et al., 2025; Asliyukse, Tonkal & Kocaoglu, 2025). These techniques are particularly effective when the spam content is textual and highly variable, as synthetic samples provide additional training patterns.

4.4 Hybrid Balancing Approaches

4.4.1 Combination of Oversampling and Undersampling

Hybrid methods combine oversampling of the minority class with undersampling of the majority class to achieve better balance while mitigating the disadvantages of each individual method (Fernández-López et al., 2018).

4.4.2 Evolutionary and Clustering-Based Balancing

Advanced hybrid approaches use clustering or evolutionary algorithms to intelligently select samples for removal or generation. Clustering helps preserve representative majority class samples, while evolutionary strategies optimize the sample distribution to maximize classifier performance. These methods have demonstrated superior performance on complex and high-dimensional datasets like Enron and SpamBase.

4.5 Recent Data-Level Strategies

4.5.1 Ensemble Sampling

Ensemble sampling generates multiple balanced datasets and trains a separate classifier on each. Predictions are combined using majority voting, which improves stability and reduces variance in performance, especially for skewed spam datasets (Feng, Zhou & Tong, 2020).

4.5.2 GAN-Based Oversampling

Generative Adversarial Networks (GANs) can synthesize realistic minority class samples, including textual spam emails, offering a more sophisticated alternative to SMOTE. GAN-based oversampling preserves complex patterns and improves model generalization, particularly when spam content contains subtle linguistic or structural cues (Akeel et al., 2025).

4.5.3 Data Augmentation for Text Data

Text-specific data augmentation techniques, such as synonym replacement, back-translation, and paraphrasing, can artificially expand the minority class. These methods generate diverse spam examples without duplicating existing samples, enhancing classifier learning for minority patterns (Asliyukse, Tonkal & Kocaoglu, 2025).

5. COST-SENSITIVE LEARNING STRATEGIES

Cost-sensitive learning strategies address class imbalance by embedding the consequences of misclassification directly into the learning process rather than altering the dataset. These approaches are particularly relevant in spam email detection, where the cost of missing a spam email (false negative) is often higher than incorrectly flagging a legitimate email (false positive) (Fernández-López et al., 2018). Cost-sensitive methods allow classifiers to prioritize

minority class accuracy while mitigating bias introduced by imbalanced distributions.

5.1 Fundamental Concepts

5.1.1 Cost Matrix and Misclassification Costs

At the core of cost-sensitive learning is the cost matrix, which specifies penalties for different types of misclassification. In a binary spam classification context, a typical matrix assigns higher costs to false negatives (spam emails classified as legitimate) than to false positives (legitimate emails classified as spam). Misclassification costs influence the learning objective of algorithms, guiding them to reduce high-cost errors even when minority class instances are scarce (Feng, Zhou & Tong, 2020). By formalizing the trade-off between sensitivity and specificity, cost-sensitive learning ensures that models focus on more critical classification errors.

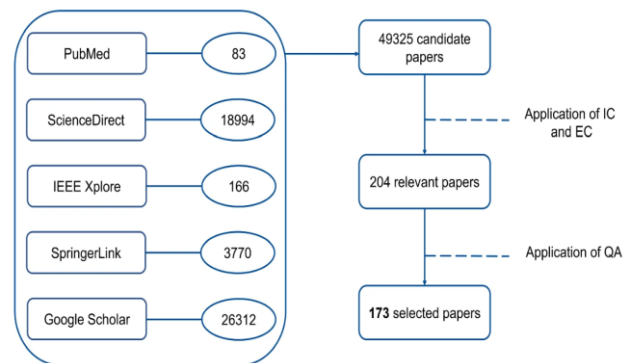


Figure-2: Cost-sensitive learning for imbalanced medical data

5.2 Algorithmic Cost Sensitivity

5.2.1 Cost-Sensitive Decision Trees

Decision tree algorithms can incorporate misclassification costs by adjusting splitting criteria or weighting leaf node predictions. Cost-sensitive decision trees prioritize splits that reduce the overall expected cost, improving minority class detection in skewed datasets (Akeel et al., 2025).

5.2.2 Cost-Sensitive SVM, kNN, Naïve Bayes

Other popular classifiers can also be adapted for cost sensitivity. In SVM, penalty parameters can be set differently for each class to minimize costly errors. k-Nearest Neighbors (kNN) can weigh neighbor votes according to misclassification costs, and Naïve Bayes can adjust prior probabilities or likelihoods to reflect cost differences (Feng, Zhou & Tong, 2020). These adaptations enhance classifier performance on minority classes without altering the dataset distribution.

5.3 Meta Costing Techniques

5.3.1 Wrapper Approaches

Meta costing refers to wrapper methods that convert any standard classifier into a cost-sensitive classifier. These methods iteratively adjust training weights or decision thresholds to minimize expected misclassification costs (Fernández-López et al., 2018). Meta costing is versatile because it allows existing algorithms to benefit from cost-sensitive optimization without modifying their internal structure, making it suitable for spam detection where multiple classifier types may be employed.

5.4 Threshold Adjustment & Calibration

5.4.1 Posterior Probability Adjustments

Cost-sensitive learning can also be applied at the decision threshold level. By adjusting posterior probabilities, classifiers can alter the probability cutoff for labeling an instance as spam or legitimate, effectively reducing the risk of high-cost errors (Akeel et al., 2025).

5.4.2 Decision Threshold Tuning

Decision thresholds can be optimized to balance sensitivity and specificity according to misclassification costs. For example, lowering the threshold for classifying spam increases recall (reduces false negatives) but may slightly reduce precision. This approach provides flexibility in real-world deployment scenarios where error costs differ.

5.5 Evaluation in Cost-Sensitive Framework

5.5.1 Cost-Weighted Metrics

Traditional metrics like accuracy are insufficient in cost-sensitive settings. Cost-weighted metrics, such as weighted F1-score or cost-sensitive error rate, incorporate misclassification penalties to evaluate performance. These metrics ensure that classifiers are assessed according to their ability to minimize costly errors rather than maximizing overall accuracy (Fernández-López et al., 2018).

5.5.2 Risk Minimization Strategies

Cost-sensitive frameworks often optimize models using risk minimization strategies, which aim to reduce the expected cost over the distribution of instances. By integrating costs into the learning objective, these strategies enhance model robustness and minority class detection, which is crucial in spam email datasets with highly skewed distributions (Feng, Zhou & Tong, 2020).

6. LITERATURE REVIEW — SYNTHESIS AND COMPARATIVE ANALYSIS

The literature on spam email classification for highly skewed datasets has evolved significantly, focusing on two primary strategies: data-level balancing and cost-sensitive learning. While both approaches aim to improve minority class detection, their methods, strengths, and weaknesses differ, as reflected in the empirical studies reviewed below.

6.1 Studies Focused on Data-Level Balancing for Spam

6.1.1 Key Papers, Datasets, Methods, and Results

Several studies have evaluated the effectiveness of data-level balancing strategies for spam detection. For instance, Akeel et al. (2025) applied SMOTE and hybrid oversampling-undersampling techniques on the UCI SpamBase and Enron datasets. They observed significant improvements in recall and F1-score for the minority (spam) class, while precision remained relatively stable. Asliyukse, Tonkal & Kocaoglu (2025) focused on LingSpam and Enron datasets, comparing SMOTE, Borderline-SMOTE, and ADASYN. Their results highlighted that synthetic oversampling enhances minority class representation and reduces bias toward majority class instances. However, they noted the potential for noise introduction in high-dimensional feature spaces. Fernández-López et al. (2018) also examined random oversampling and undersampling, emphasizing the trade-off between reducing bias and preserving information, which aligns with findings from recent hybrid strategies. Collectively, these studies suggest that data-level balancing is highly effective in improving spam recall but may require careful tuning to avoid overfitting or information loss.

6.2 Studies Focused on Cost-Sensitive Learning

6.2.1 Key Works and Performance vs. Imbalance

Cost-sensitive learning studies aim to integrate misclassification costs into learning algorithms to prioritize minority class detection. Feng, Zhou & Tong (2020) investigated cost-sensitive SVM and decision trees on SpamBase and Enron datasets, reporting improved minority-class recall without duplicating samples. Akeel et al. (2025) demonstrated that cost-sensitive Naïve Bayes and kNN also increased the detection of rare spam instances while maintaining acceptable precision. These methods are particularly beneficial when dataset augmentation is not feasible, as they preserve the natural distribution of emails. Limitations noted across studies include the difficulty in selecting appropriate cost matrices and the sensitivity of results to threshold settings, which may affect generalization across different spam datasets.

6.3 Direct Comparative Studies

6.3.1 Comparative Findings, Strengths, and Weaknesses

Direct comparisons of data-level and cost-sensitive approaches are limited but insightful. Asliyukse, Tonkal & Kocaoglu (2025) compared SMOTE-based oversampling against cost-sensitive SVM on the LingSpam dataset. They found that synthetic oversampling produced higher recall, while cost-sensitive methods provided better precision-stability, particularly when spam prevalence was extremely low. Hybrid approaches that combine oversampling with cost-sensitive learning were shown to outperform either strategy alone, achieving a balance between recall and precision across diverse spam datasets (Fernández-López et al., 2018). Weaknesses in existing comparative studies include the lack of standardized evaluation metrics and limited exploration across multiple, heterogeneous spam datasets.

6.4 Trends in Research

6.4.1 Dominant Approaches and Emerging Directions

Current trends indicate that hybrid frameworks integrating both data-level balancing and cost-sensitive learning are becoming dominant, as they leverage the strengths of each strategy while mitigating weaknesses. Deep learning approaches, combined with text augmentation and cost-sensitive loss functions, are emerging as powerful alternatives for high-dimensional, evolving spam datasets (Akeel et al., 2025). Additionally, researchers are increasingly using ensemble methods to stabilize predictions across imbalanced samples and improve robustness. These trends point toward multi-strategy pipelines that can adapt to dynamic spam patterns and maintain performance across varying dataset characteristics.

7. CRITICAL DISCUSSION

The critical discussion synthesizes insights from the literature on data-level balancing and cost-sensitive learning strategies for imbalanced spam datasets. It evaluates their relative effectiveness, the influence of dataset characteristics, practical relevance, and limitations observed in current research.

7.1 Effectiveness Comparison

7.1.1 One Family of Approach Outperforms the Other

Data-level balancing techniques, particularly synthetic oversampling methods like SMOTE and ADASYN, generally perform well in improving recall for the minority (spam) class by increasing sample representation and diversity (Akeel et al., 2025). They are especially effective when the dataset contains sufficient features to generate meaningful synthetic instances. In contrast, cost-sensitive learning tends

to outperform data-level methods in scenarios where dataset augmentation is infeasible or when overfitting is a concern, as it preserves the original data distribution while emphasizing costly misclassifications (Feng, Zhou & Tong, 2020). Hybrid approaches that combine both methods have shown the most consistent performance, achieving balanced recall and precision across different spam datasets. These findings suggest that the choice of method depends on dataset size, dimensionality, and the acceptable trade-off between false positives and false negatives.

7.2 Impact of Dataset Characteristics

7.2.1 Influence of Imbalance Ratio, Feature Sparsity, and Noise

The effectiveness of both data-level and cost-sensitive strategies is significantly influenced by dataset characteristics. High imbalance ratios exacerbate bias toward the majority class, often requiring more aggressive resampling or cost adjustments to maintain minority class detection (Fernández-López et al., 2018). Feature sparsity, common in text-based spam datasets with high-dimensional n-gram representations, can reduce the ability of oversampling to generate informative synthetic samples, and may necessitate dimensionality reduction or feature selection. Noise and outliers in email datasets can further complicate model learning, as both oversampling and cost-sensitive methods may amplify errors if noise is not filtered, highlighting the need for preprocessing and careful algorithmic tuning.

7.3 Practical Implications for Spam Detection Systems

7.3.1 Industrial Relevance

From an industrial perspective, imbalanced learning strategies are essential for deploying reliable spam detection systems. False negatives (missed spam) carry risks such as phishing attacks, malware propagation, and reputational damage, making minority-class performance a critical metric (Akeel et al., 2025). Data-level methods can be integrated into batch training pipelines, while cost-sensitive approaches allow real-time classification without modifying the dataset, which is advantageous in large-scale email systems. Hybrid frameworks and ensemble strategies provide further robustness, supporting deployment in dynamic, production-level environments where spam patterns continuously evolve.

7.4 Limitations in Existing Research

7.4.1 Lack of Standardized Benchmarking

A major limitation in the current literature is the absence of standardized benchmarking, making it difficult to compare studies objectively. Different papers use varied datasets,

preprocessing steps, and evaluation metrics, limiting reproducibility and generalizability (Fernández-López et al., 2018).

7.4.2 Limited Comparisons Across Multiple Datasets

Many studies focus on a single dataset, such as SpamBase or Enron, without cross-dataset validation. This restricts understanding of how methods perform under varying characteristics, such as imbalance ratio, feature space, or email content diversity (Akeel et al., 2025).

7.4.3 Neglect of Evolving Spam Patterns

Finally, existing research often assumes static spam distributions. In practice, spammers continuously evolve their tactics, introducing new keywords, obfuscation techniques, and email structures. Few studies examine how imbalance strategies perform under concept drift or dynamically changing spam patterns, which is critical for designing adaptive spam detection systems (Asliyukse, Tonkal & Kocaoglu, 2025).

8. CONCLUSION

This review critically examined strategies for handling class imbalance in highly skewed spam email datasets, focusing on data-level balancing and cost-sensitive learning approaches. Data-level techniques, including random oversampling, undersampling, SMOTE variants, and hybrid methods, have been widely employed to enhance minority class representation. These approaches generally improve recall and F1-score for spam detection by increasing the number of training examples or optimizing class distribution. However, they are susceptible to overfitting, noise introduction, and information loss in high-dimensional feature spaces. Cost-sensitive learning strategies, encompassing cost-adjusted decision trees, SVM, kNN, Naïve Bayes, and meta-cost frameworks, provide an alternative by embedding misclassification penalties directly into the learning process. These methods preserve the original dataset distribution while emphasizing errors on minority instances, improving robustness in scenarios where resampling is infeasible. Comparative studies indicate that neither strategy alone is universally superior; hybrid and ensemble approaches that integrate data-level and cost-sensitive frameworks tend to achieve more balanced performance, effectively optimizing both precision and recall across diverse datasets. The literature also highlights the critical influence of dataset characteristics, such as imbalance ratio, feature sparsity, and noise, on classifier performance, emphasizing the need for tailored solutions. Emerging research trends include deep learning models combined with cost-sensitive loss functions, advanced text augmentation, and GAN-based oversampling. Overall, the synthesis underscores the importance of method selection based on dataset properties, application requirements, and practical constraints, providing a comprehensive roadmap for researchers and practitioners

seeking to enhance spam detection in imbalanced email environments.

8.1. Limitations of the Review

Despite offering a comprehensive synthesis of data-level and cost-sensitive approaches for spam email classification, this review has certain limitations. First, the analysis relies primarily on published studies using benchmark datasets such as SpamBase, Enron, and LingSpam, which may not capture the full diversity of real-world email traffic. Second, the review is constrained by the availability of direct comparative studies between data-level and cost-sensitive methods, limiting the ability to generalize performance conclusions across all imbalanced scenarios. Third, rapidly evolving spamming techniques and concept drift are not fully addressed in the existing literature, meaning the effectiveness of reviewed strategies may differ in dynamic operational environments. Finally, emerging methods such as transformer-based models and GAN-generated data are still in early stages, and their long-term efficacy remains underexplored.

REFERENCES

1. Akeel A., Butt K.K., Javed K., Tariq M. & Yousaf M. (2025) Email Spam Detection Using Machine Learning with Optimized Feature Engineering and Classification Techniques. *Journal of Computing & Biomedical Informatics*.
2. Asliyukse H., Tonkal Ö. & Kocaoglu R. (2025) A Comparative Evaluation of a Multimodal Approach for Spam Email Classification Using DistilBERT and Structural Features. *MDPI*.
3. Fernández-López M., García S., Herrera F., Chawla N.V. & Krawczyk B. (2018) 'Strategies to mitigate class imbalance' in *Artificial Intelligence Review*, Springer, pp. 1–38.
4. Feng Y., Zhou M. & Tong X. (2020) Imbalanced classification: a paradigm-based review. *arXiv preprint arXiv:2005.04542*.
5. Jáñez-Martino F., Alaiz-Rodríguez R., González-Castro V., Fidalgo E. & Alegre E. (2022) A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*.
6. Gangavarapu T., Jaidhar C.D. & Chanduka B. (2020) Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review*.
7. Shaukat A. et al. (2024) Improving spam email classification accuracy using ensemble techniques: a

- stacking approach. International Journal of Information Security.
8. Mishra G. & Gautam P. (2025) Improving Email Spam Detection and Classification Through Data Balancing and Ensemble Machine Learning-Based Boosting Approaches. International Journal of Environmental Sciences.
 9. Spam Detection in Emails Using Machine Learning Techniques (2024) IJCIIT.
 10. Shirvani G. & Ghasemshirazi S. (2025) Advancing Email Spam Detection: Leveraging Zero-Shot Learning and Large Language Models. arXiv preprint.
 11. A survey on imbalanced learning: latest research, applications and challenges (2024) Springer.
 12. Feng Y., Zhou M. & Tong X. (2020) Imbalanced classification: a paradigm-based review. Wiley.
 13. A Systematic Review on Imbalanced Data Challenges in Machine Learning (2020) ACM.
 14. Ratadiya P. & Moorthy R. (2019) Spam filtering on forums: Synthetic oversampling for imbalanced classification. arXiv preprint.
 15. Cost-Sensitive Machine Learning (Encyclopedia entry) by Ling & Sheng.
 16. Arora J. et al. (2022) MCBC-SMOTE: Majority Clustering for Balanced Classification. Computers, Materials & Continua.
 17. Chawla N.V., Bowyer K.W., Hall L.O. & Kegelmeyer W.P. (2011) SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research.
 18. Singh M. et al. (2020) Resolving the imbalance issue in short messaging service spam dataset using cost-sensitive techniques. Journal of Information Security and Applications.
 19. Zhao C., Xin Y., Li X. et al. (2020) Heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. Applied Sciences.
 20. Cost-Sensitive Learning for Imbalanced Classification: A Comprehensive Review (2024) Artificial Intelligence Review.
 21. Cost-Sensitive Deep Learning and Ensemble Methods (CSE-IDS 2021) Computers & Security.
 22. Meta-Cost: A General Method for Making Classifiers Cost-Sensitive (Original Cost-Sensitive Learning paper, 1999) – foundational method referenced widely.
 23. A Comparative Analysis of Classical Machine Learning Algorithms for Spam Detection (IJCT 2025).
 24. Classification Model of Spam Emails Based on Data Mining – Deep Learning Techniques (IETA 2023).
 25. Improving Knowledge Based Spam Detection Methods: The Effect of Feature Engineering (SCIRP, 2017).
 26. Alkhodhairy G. & Saleem K. (2025) Machine learning algorithm for detecting suspicious email messages using NLP. Alexandria Engineering Journal.
 27. Advances in spam detection for email spam, web spam, social spam (SAGE 2021).
 28. Hybrid Clustering Strategies for Effective Oversampling and Undersampling (Nature Scientific Reports 2024) – generalized imbalance technique.
 29. He H. & Garcia E. (2009) Learning from Imbalanced Data. IEEE Transactions.
 30. Imbalance in Machine Learning: Foundations, Algorithms, Applications (Book 2013).
 31. Sajid A. et al. (2017) LIUBoost: Locality Informed Underboosting for Imbalanced Data. arXiv.
 32. Shu R. et al. (2022) Reducing the Cost of Training Security Classifier via Optimized Semi-Supervised Learning. arXiv.
 33. Deep Learning based Frameworks for Handling Imbalance in Email and URL Data Analysis (Simran K. et al. 2020).
 34. A Stacking Approach Machine Learning for Spam Email Detection (2025, PMC).
 35. Spam vs. Ham NLP classifier – Feature Engineering & Resampling (Case discussion, methodological insight).