

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 - Spam email detection has become a critical research area due to the increasing volume of unsolicited emails and their role in phishing and malware delivery. A major challenge in spam classification is the presence of highly imbalanced datasets, where legitimate emails dominate spam samples, leading to biased learning behavior in traditional machine learning classifiers. This study presents a comparative analysis of data-level balancing strategies and cost-sensitive learning methods for spam classification under extreme skew conditions. Three benchmark datasets, namely SpamAssassin, Ling-Spam, and CSDMC2010, are utilized and further modified to generate highly skewed variants with imbalance ratios of 90:10, 95:5, and 98:2. Text preprocessing is applied, followed by TF-IDF feature extraction. Experiments are performed using Multinomial Naïve Bayes, Support Vector Machine, and Random Forest classifiers. Data-level strategies such as Random Undersampling (RUS), Random Oversampling (ROS), and SMOTE are compared against cost-sensitive techniques including class-weighted SVM, MetaCost, and decision threshold tuning. Results demonstrate that SMOTE combined with Random Forest yields strong improvements in recall and F2-score, whereas threshold-tuned SVM achieves better precision and reduces false positives. The findings confirm that cost-sensitive learning offers stable performance under severe skewness, while resampling methods provide significant recall gains but may increase computational overhead. This study contributes a structured evaluation framework for selecting imbalance-handling techniques in real-world spam filtering systems.

Key Words: Spam Detection, Class Imbalance, Cost-Sensitive Learning, SMOTE, Threshold Tuning, TF-IDF, Imbalanced Classification

1. INTRODUCTION

Spam email detection remains a crucial research domain in cybersecurity and machine learning due to the rapid growth of electronic communication systems. Email is widely used for professional communication, academic correspondence, e-commerce transactions, and official notifications, making it an essential component of modern digital infrastructure. However, this popularity also attracts malicious actors who exploit email as a low-cost and high-reach medium to distribute unsolicited and harmful messages. Spam emails

not only cause inconvenience by overwhelming inboxes but also create severe security threats by serving as a carrier for phishing attacks, identity theft, and malware distribution. As spam techniques continue to evolve, there is an increasing need for robust and intelligent spam filtering mechanisms that can adapt to changing attack strategies and effectively detect malicious content.

1.1 Background

Email communication has become a foundational tool for global interaction, supporting both organizational operations and personal communication. With the expansion of internet connectivity and cloud-based messaging platforms, email has achieved widespread adoption because of its efficiency, asynchronous communication model, and ability to transfer large-scale information instantly. Despite its advantages, email is frequently targeted by spammers due to its open architecture and massive user base. Spam emails represent unsolicited bulk messages that are typically sent for advertising, scams, or malicious purposes. The growing volume of spam has made automated spam filtering an essential requirement for ensuring secure and productive email communication (He and Garcia, 2009).

1.1.1 Importance of Email Communication and Spam Threats

Email is considered one of the most reliable digital communication tools due to its formal nature, accessibility, and ability to support secure document exchange. Governments, businesses, educational institutions, and individuals rely heavily on email for official notifications, legal communication, and financial transactions. However, spam emails have become a major challenge, contributing to wasted bandwidth, reduced productivity, and increased cybersecurity risks. Many spam messages are designed to manipulate recipients into revealing confidential information, downloading malicious attachments, or visiting fraudulent websites. Consequently, spam filtering is not only a matter of convenience but also a significant cybersecurity requirement in modern information systems (Metsis, Androutsopoulos and Paliouras, 2006).

1.2 Problem of Highly Skewed Spam Datasets

A major technical challenge in spam email classification is the highly imbalanced nature of real-world datasets. In most practical environments, legitimate emails (ham) significantly outnumber spam emails. This imbalance creates biased learning conditions where machine learning models tend to prioritize the majority class. As a result, classifiers may achieve high overall accuracy but fail to detect spam emails effectively. This issue becomes even more severe in corporate and enterprise environments, where spam may represent only a very small fraction of the overall email traffic. Therefore, handling dataset skewness is essential for designing effective spam detection systems (Japkowicz and Shah, 2011).

1.2.1 Imbalance Ratios such as 95:5 and 98:2

Dataset imbalance is often expressed as a ratio of majority class samples to minority class samples. In spam detection, a dataset with a 95:5 ratio means that 95% of emails are ham while only 5% are spam. Similarly, a 98:2 ratio represents an extreme scenario where spam emails constitute only 2% of the total dataset. Such extreme skewness causes classifiers to become overly conservative in predicting spam because predicting ham most of the time results in fewer classification errors overall. Consequently, the model may fail to identify rare spam emails, making it ineffective for real-world deployment. These imbalance conditions require specialized strategies such as resampling methods or cost-sensitive learning approaches to improve spam detection performance (Chawla et al., 2002).

1.3 Cost Asymmetry in Spam Filtering

Spam filtering is not only affected by class imbalance but also by cost asymmetry, meaning that different types of misclassification errors have unequal consequences. A false positive occurs when a legitimate email is incorrectly classified as spam, while a false negative occurs when a spam email is incorrectly classified as legitimate. These errors have different impacts on users and organizations. In real-world email systems, false positives are generally considered more critical because they may lead to missed business opportunities, loss of legal documents, or failure to receive important personal communication. Therefore, spam filtering systems must be designed to minimize such costly errors (Drummond and Holte, 2006).

1.3.1 False Positives vs False Negatives Impact

False positives can have serious consequences, particularly in corporate and institutional environments. If an important email such as a job offer, legal notice, or financial transaction confirmation is incorrectly blocked, it may result in significant harm. On the other hand, false negatives allow spam messages to reach the inbox, causing annoyance and potential exposure to phishing or malware. While both

errors are undesirable, the cost of false positives is often higher because legitimate communication may be permanently lost. This makes spam filtering fundamentally different from many other classification problems, as the objective is not only high accuracy but also balanced decision-making based on real-world consequences (Powers, 2011).

1.4 Research Objectives

The primary objective of this research is to conduct a systematic comparison of data-level balancing strategies and cost-sensitive learning techniques for highly skewed spam email datasets. The study aims to determine which approach provides the most reliable performance under increasing skewness while maintaining acceptable computational efficiency. Since spam detection systems must operate in real-time and handle large-scale datasets, identifying a method that balances accuracy, recall, precision, and computational cost is essential.

1.4.1 Comparative Analysis of Data-Level vs Cost-Sensitive Approaches

This research evaluates data-level methods such as Random Undersampling, Random Oversampling, and SMOTE, which modify the dataset distribution to improve minority class learning. These are compared with cost-sensitive methods such as class-weighted learning, MetaCost, and threshold tuning, which modify the classifier's decision-making process. The goal is to quantify how each strategy impacts spam detection performance, particularly under severe imbalance ratios (Chawla et al., 2002).

1.4.2 Identifying the Most Robust Strategy Under Increasing Skewness

A major objective of this research is to identify which technique remains stable when imbalance increases from moderate levels to extreme skewness. This involves evaluating classifier performance across multiple dataset configurations such as 90:10, 95:5, and 98:2. The study focuses on determining whether data-level balancing methods remain effective or whether cost-sensitive strategies provide better stability under extreme conditions (Saito and Rehmsmeier, 2015).

2. RELATED WORK

Spam filtering has been widely studied as a classical text classification problem in machine learning and natural language processing. Over the past two decades, researchers have proposed various statistical and computational techniques to improve spam detection accuracy, reduce false positives, and enhance model robustness against evolving spam strategies. However, the presence of highly imbalanced datasets remains a major challenge, particularly because real-world email environments often contain far fewer spam

messages compared to legitimate emails. Consequently, modern research increasingly focuses on imbalance-handling strategies such as resampling, synthetic data generation, and cost-sensitive learning approaches. This section reviews existing literature on spam filtering algorithms, class imbalance challenges, balancing strategies, cost-sensitive methods, and evaluation metrics relevant to highly skewed spam datasets.

2.1 Spam Filtering and Traditional ML Approaches

Spam email detection has traditionally been addressed using supervised machine learning models trained on labeled email datasets. Early machine learning-based spam filters replaced rule-based approaches by learning statistical patterns directly from email content. Among these, Naïve Bayes classifiers became one of the earliest and most widely adopted methods due to their simplicity, probabilistic foundation, and strong performance in high-dimensional text spaces. The success of Naïve Bayes in spam filtering is largely attributed to its ability to efficiently compute class probabilities even when the number of features is extremely large, such as when Bag-of-Words or TF-IDF representations are used (Metsis, Androutsopoulos and Paliouras, 2006).

2.2 Class Imbalance in Text Classification

Class imbalance refers to a dataset condition where one class contains significantly more samples than the other. In spam filtering, legitimate emails (ham) typically dominate the dataset, while spam messages form a minority class. This imbalance negatively affects classifier training because most machine learning algorithms are designed to minimize overall error. As a result, classifiers may become biased toward the majority class and fail to detect minority class instances effectively. In highly skewed spam datasets, a classifier may incorrectly classify most spam messages as ham while still achieving high accuracy due to the overwhelming majority of legitimate emails (He and Garcia, 2009).

2.3 Data-Level Balancing Techniques

Data-level balancing strategies attempt to reduce class imbalance by modifying the dataset distribution before model training. These techniques aim to either increase the number of minority class samples or reduce the number of majority class samples so that the classifier can learn spam patterns more effectively. Data-level methods are widely used because they are classifier-independent and can be applied as preprocessing steps without changing the learning algorithm itself. The most common balancing techniques include Random Undersampling (RUS), Random Oversampling (ROS), and synthetic sampling methods such as SMOTE (Chawla et al., 2002).

2.3.1 Random Undersampling (RUS)

Random Undersampling is a technique that reduces the majority class size by randomly removing ham emails until a more balanced class distribution is achieved. This approach is computationally efficient because it reduces the dataset size, leading to faster training time. However, the main disadvantage of RUS is that it may discard important ham samples that contain useful information for classification. Removing such samples can reduce the classifier's ability to distinguish between legitimate and spam messages, potentially increasing false positives. Despite this limitation, RUS is often used in large-scale applications where computational efficiency is critical (He and Garcia, 2009).

2.4 Cost-Sensitive Learning Techniques

Unlike data-level balancing methods, cost-sensitive learning approaches handle imbalance by modifying the learning algorithm rather than the dataset distribution. These methods assign higher penalties to misclassifications involving the minority class or critical errors such as false positives. Cost-sensitive approaches are especially important in spam filtering because false positives (ham classified as spam) may cause serious consequences such as missed business communication. Cost-sensitive learning allows classifiers to incorporate real-world misclassification costs, making them more practical for deployment in real email filtering systems (Elkan, 2001).

2.4.1 Class Weighting in SVM

Class weighting is a commonly used cost-sensitive technique in Support Vector Machines. In this approach, the optimization function assigns higher penalty weights to minority class errors. This forces the classifier to focus more on correctly detecting spam messages. Class-weighted SVM is effective because it does not require altering the dataset distribution and works naturally within margin-based learning frameworks. By increasing the cost of misclassifying spam emails, the classifier shifts its decision boundary toward the majority class, improving recall for spam detection (Veropoulos, Campbell and Cristianini, 1999).

2.5 Evaluation Metrics for Imbalanced Spam Classification

Evaluation of spam detection models under imbalance conditions requires careful selection of performance metrics. Accuracy is often misleading because it can remain high even when the classifier fails to detect spam. Therefore, researchers emphasize confusion-matrix-based metrics such as precision, recall, and F-scores. Precision measures how many predicted spam emails are actually spam, reflecting the false positive rate. Recall measures how many actual spam emails are correctly identified, representing detection capability. The F1-score provides a harmonic balance between precision and recall, while the F2-score gives

greater weight to recall, making it more suitable when detecting spam is more critical than avoiding false alarms (Powers, 2011).

3. MATERIALS AND METHODS

This section presents the methodology adopted for conducting a comparative study of cost-sensitive learning and data-level balancing strategies on highly skewed spam email datasets. The overall experimental design follows a structured workflow involving dataset preparation, preprocessing, feature extraction, imbalance-handling, model training, and performance evaluation. Since spam email classification is a text mining task, the research uses standard natural language processing (NLP) techniques combined with machine learning classifiers. Additionally, the study emphasizes imbalance-aware learning by simulating real-world skewness conditions and evaluating multiple strategies to address minority class under-representation (He and Garcia, 2009).

3.1 Research Workflow

The research workflow is designed as a systematic pipeline to ensure fair comparison among different imbalance-handling techniques. The workflow begins with dataset collection, followed by text preprocessing to remove noise and standardize email content. After preprocessing, the textual emails are converted into numerical representations using TF-IDF vectorization, which transforms the email text into high-dimensional feature vectors suitable for machine learning. Once features are extracted, imbalance-handling strategies are applied in two categories: data-level balancing methods, such as oversampling and undersampling, and cost-sensitive learning techniques, such as class weighting and threshold tuning. After applying these strategies, machine learning classifiers are trained using the processed training dataset. Finally, model evaluation is conducted using imbalance-aware performance metrics such as precision, recall, F1-score, F2-score, and AUC-PR to ensure reliable assessment under skewed distributions (Saito and Rehmsmeier, 2015).

3.1.1 Pipeline: Preprocessing → TF-IDF → Balancing/Cost-Sensitive → Training → Evaluation

The experimental pipeline follows a sequential process where each stage contributes to improving classification reliability. First, preprocessing ensures that irrelevant elements such as HTML tags, punctuation, and stopwords do not distort the feature space. Next, TF-IDF converts the cleaned text into weighted vectors representing term importance. After feature transformation, balancing methods such as SMOTE or undersampling are applied to the training data to improve spam representation, while cost-sensitive methods modify the learning behavior by assigning different penalties to misclassification errors. Following this, classifiers are trained on the prepared dataset. Finally,

evaluation is conducted on the test set using suitable metrics, ensuring that results are not biased by the imbalance problem and that minority class performance is accurately captured (Chawla et al., 2002).

3.2 Dataset Description

This research uses three publicly available benchmark datasets to ensure generalizability and reproducibility. The datasets include SpamAssassin, Ling-Spam, and CSDMC2010, which are widely used in spam filtering research. These datasets contain labeled email messages categorized into spam and ham. The use of multiple datasets is important because spam content varies across sources, and a model that performs well on one dataset may not generalize well to others. Moreover, using multiple datasets helps evaluate the stability of imbalance-handling techniques across diverse spam distributions and linguistic patterns (Metsis, Androutsopoulos and Paliouras, 2006).

3.2.1 SpamAssassin Dataset

The SpamAssassin Public Corpus is one of the most widely used datasets for spam email research. It contains real-world spam and legitimate emails collected from multiple sources. The dataset includes a mixture of commercial spam messages, phishing-like emails, and normal personal or professional emails. SpamAssassin is valuable for experimentation because it contains realistic spam characteristics such as embedded URLs, advertising content, and obfuscated words. The dataset is frequently used for benchmarking machine learning spam classifiers due to its open accessibility and reliable labeling (Metsis, Androutsopoulos and Paliouras, 2006).

3.3 Creation of Highly Skewed Dataset Variants

To simulate real-world spam filtering conditions, highly skewed dataset variants were created from each benchmark dataset. Many real email environments contain a very small proportion of spam compared to legitimate emails, often reaching imbalance ratios such as 95:5 or even 98:2. Therefore, this research generated controlled dataset variants with imbalance ratios of 90:10, 95:5, and 98:2. This was achieved by randomly reducing the number of spam samples while keeping the ham class unchanged, ensuring that the dataset size remains realistic and the imbalance is systematically increased. Such dataset simulation allows evaluation of classifier performance under increasing skewness and provides insight into which imbalance-handling strategies remain stable when spam becomes extremely rare (He and Garcia, 2009).

3.4 Text Preprocessing

Preprocessing is a crucial step in spam email classification because raw email data often contains noise such as punctuation, HTML tags, URLs, and irrelevant tokens. If such

elements are not removed, they can distort feature representation and reduce classifier performance. This research applied a standard NLP preprocessing pipeline to normalize the email text and reduce vocabulary size. Preprocessing also helps improve the effectiveness of TF-IDF feature extraction by ensuring that only meaningful textual components are included in the feature space. Proper preprocessing enhances model generalization by reducing over fitting to irrelevant email formatting patterns (Manning, Raghavan and Schütze, 2008).

3.4.1 Tokenization, Stop word Removal, and Stemming/Lemmatization

Tokenization is the process of splitting text into individual words or tokens. After tokenization, stopwords such as “the”, “is”, “and”, and “of” are removed because they occur frequently but provide minimal discriminatory information for spam classification. Following this, stemming or lemmatization is applied to reduce words to their root forms. For instance, “running”, “runs”, and “ran” may be reduced to “run”, improving vocabulary consistency. These preprocessing steps reduce feature dimensionality and help classifiers learn meaningful spam patterns more effectively (Manning, Raghavan and Schütze, 2008).

3.5 Feature Extraction

Machine learning classifiers require numerical input, but email messages are inherently unstructured text. Therefore, feature extraction is required to convert emails into numerical vectors. In this study, TF-IDF vectorization was used because it effectively represents text by capturing term importance relative to the document corpus. TF-IDF is widely used in spam filtering because spam emails often contain distinct keywords and repeated promotional phrases that receive higher weights in TF-IDF representation (Metsis, Androutsopoulos and Paliouras, 2006).

3.6 Base Classifiers Used

To ensure a comprehensive comparison, this study employed three diverse machine learning classifiers: Multinomial Naïve Bayes, Support Vector Machine, and Random Forest. The selection of these classifiers was motivated by their strong performance in text classification tasks and their distinct learning principles. Using classifiers with different learning behaviors provides a fair evaluation of whether imbalance-handling techniques generalize across algorithms. Additionally, these models represent probabilistic, margin-based, and ensemble-based learning paradigms, making them suitable for comparative analysis in spam filtering research (Breiman, 2001).

4. IMBALANCE HANDLING STRATEGIES

Class imbalance is a dominant challenge in spam email classification because legitimate emails usually far

outnumber spam emails in real-world inbox traffic. This imbalance causes traditional machine learning models to become biased toward the majority class, leading to poor spam detection performance, especially under extreme skewness such as 95:5 or 98:2 distributions. To address this issue, imbalance-handling strategies are generally categorized into two major groups: data-level balancing methods, which modify the dataset distribution, and cost-sensitive learning methods, which modify the classifier’s learning objective. This study evaluates both categories to identify the most effective approach for improving spam detection under highly imbalanced conditions (He and Garcia, 2009).

4.1 Data-Level Balancing Methods

Data-level balancing strategies attempt to improve minority class learning by altering the dataset composition before training. These methods either increase minority class samples or reduce majority class samples to create a more balanced distribution. The primary advantage of data-level approaches is that they are independent of the classifier, meaning they can be applied to any machine learning algorithm. However, these methods may introduce limitations such as information loss or overfitting, depending on the chosen balancing technique. In spam filtering, data-level balancing is particularly useful because it allows the classifier to observe more spam patterns during training, improving its ability to generalize minority class behavior (Chawla et al., 2002).

4.2 Cost-Sensitive Learning Methods

Cost-sensitive learning methods handle imbalance by modifying the classifier training objective rather than changing the dataset distribution. These approaches incorporate misclassification costs into learning, ensuring that errors on minority class instances or high-cost errors are penalized more heavily. Cost-sensitive learning is particularly important in spam filtering because false positives, where legitimate emails are wrongly blocked, can be more damaging than false negatives. By introducing cost sensitivity, classifiers can be trained to reflect real-world priorities, improving practical usability. Unlike resampling methods, cost-sensitive approaches typically avoid artificial changes to dataset distribution and may reduce the risk of overfitting caused by oversampling (Elkan, 2001).

4.2.1 Class-Weighted SVM

Class-weighted Support Vector Machine is a cost-sensitive method where different penalty weights are assigned to the spam and ham classes during training. In this approach, the optimization function applies a higher penalty when the minority class (spam) is misclassified, forcing the classifier to shift the decision boundary in favor of spam detection. The key advantage of class-weighted SVM is that it improves recall for spam without altering the dataset composition.

This makes it computationally efficient and suitable for large-scale spam filtering systems. However, the limitation is that assigning excessively high weights may lead to an increase in false positives, as the classifier becomes overly aggressive in labeling emails as spam. Therefore, weight tuning is essential to achieve an appropriate balance between precision and recall (Veropoulos, Campbell and Cristianini, 1999).

5. EXPERIMENTAL SETUP

The experimental setup is designed to ensure a controlled and fair comparison between different imbalance-handling strategies. Since the objective of this study is comparative analysis, it is essential that all models are evaluated under identical conditions, including consistent datasets, preprocessing steps, feature extraction, and classifiers. This controlled experimental framework ensures that differences in results can be attributed to the imbalance-handling methods rather than variations in data preparation or model configuration. Furthermore, evaluation is performed using imbalance-aware metrics to provide meaningful insights into spam detection performance under skewed distributions (Saito and Rehmsmeier, 2015).

5.1 Experimental Design

The study follows a controlled comparison methodology where the same datasets and the same feature extraction approach are used across all experiments. The base classifiers, namely Naïve Bayes, SVM, and Random Forest, are trained under multiple conditions: baseline (no balancing), data-level balancing (RUS, ROS, SMOTE), and cost-sensitive learning (class weighting, MetaCost, threshold tuning). By keeping the experimental pipeline constant and varying only the imbalance-handling strategy, the study ensures that performance differences are fairly evaluated. This design provides a clear understanding of which method offers the best robustness under increasing skewness conditions (He and Garcia, 2009).

5.2 Train-Test Split and Cross Validation

To ensure reliable performance evaluation, the dataset is divided into training and testing subsets using stratified splitting. Stratification ensures that the proportion of spam and ham emails remains consistent across training and testing datasets. This is important in imbalanced learning because random splitting may create subsets with very few spam samples, resulting in unreliable evaluation. In addition, cross-validation is applied to ensure that results are stable and not dependent on a particular split. A critical aspect of this setup is avoiding data leakage: balancing techniques such as SMOTE and oversampling are applied only on the training dataset, while the test dataset remains untouched. This prevents artificial inflation of performance metrics and ensures realistic model evaluation (Japkowicz and Shah, 2011).

5.3 Hyper parameter Tuning

Hyperparameter tuning is performed to optimize classifier performance and ensure fair comparison. This study uses GridSearchCV to explore multiple hyperparameter combinations systematically. For Naïve Bayes, smoothing parameter alpha is tuned to control probability estimation stability. For SVM, the penalty parameter C is optimized, along with kernel selection if required. For Random Forest, key parameters such as the number of trees (n_estimators), maximum tree depth (max_depth), and minimum samples split are tuned. GridSearchCV is performed using cross-validation on the training set, ensuring that the selected hyperparameters generalize effectively. Hyperparameter tuning is essential because imbalance-handling strategies may influence optimal model settings differently, especially in highly skewed datasets (Bergstra and Bengio, 2012).

5.4 Evaluation Metrics

Evaluation metrics play a central role in imbalanced spam classification because traditional accuracy fails to capture minority class detection capability. Therefore, multiple metrics are used to evaluate classifier performance comprehensively. The metrics include confusion matrix, precision, recall, F1-score, F2-score, and AUC-PR. Each metric provides a different perspective on spam detection performance, ensuring a balanced evaluation framework. In particular, recall and F2-score are emphasized because missing spam emails can allow phishing and malware attacks to reach users (Powers, 2011).

5.4.1 Confusion Matrix

The confusion matrix is used to summarize classification outcomes by reporting true positives, true negatives, false positives, and false negatives. In spam filtering, true positives represent correctly detected spam emails, while false positives represent legitimate emails wrongly labeled as spam. The confusion matrix is essential because it provides a clear understanding of error distribution, which is critical in cost-sensitive spam filtering environments. It also forms the basis for calculating other evaluation metrics such as precision and recall (Fawcett, 2006).

5.5 Computational Environment

All experiments are conducted using Python programming language in a standard machine learning environment. The implementation uses widely adopted libraries such as NumPy and Pandas for data handling, Scikit-learn for machine learning models and evaluation, and Imbalanced-learn for resampling techniques such as SMOTE. The computational setup includes a system with at least an Intel i5/i7 processor (or equivalent), 8GB or higher RAM, and sufficient storage for dataset processing. Since TF-IDF feature matrices are sparse and high-dimensional, memory efficiency is an important consideration.

6. RESULTS

This section presents a detailed analysis of experimental findings obtained from applying baseline classifiers, data-level balancing methods, and cost-sensitive learning strategies to highly skewed spam email datasets. The evaluation focuses on imbalance-aware performance metrics, particularly recall, F2-score, precision, and AUC-PR. Results are analyzed across different skewness levels (90:10, 95:5, and 98:2) to assess method robustness.

6.1 Dataset Skewness Statistics

Before applying any imbalance-handling strategy, the datasets were analyzed to confirm the distribution of ham and spam emails under different simulated skew conditions. The imbalance ratio significantly influences classifier behavior, particularly under extreme skewness such as 98:2. As the proportion of spam decreases, minority class detection becomes more challenging due to limited representation during training (He and Garcia, 2009).

6.2 Baseline Results Without Any Strategy

Baseline experiments were conducted without applying any balancing or cost-sensitive techniques. The results demonstrate that classifier performance deteriorates significantly as imbalance increases, especially at the 98:2 ratio. Although overall accuracy remains high, recall and F2-score decline sharply, indicating failure to detect spam effectively.

6.2.1 Performance Collapse Under 98:2 Imbalance

Under extreme skewness (98:2), classifiers such as Naïve Bayes and SVM predict the majority class in most cases. While accuracy may exceed 97%, recall drops substantially because many spam emails are misclassified as ham. This behavior confirms that traditional classifiers are not inherently robust to severe imbalance conditions (Japkowicz and Shah, 2011).

6.2.2 Accuracy Paradox Highlight

The baseline results clearly demonstrate the accuracy paradox, where high accuracy masks poor minority class performance. For example, a classifier predicting all emails as ham achieves 98% accuracy in a 98:2 dataset but fails entirely at spam detection. This reinforces the importance of using recall, F2-score, and AUC-PR rather than accuracy alone when evaluating spam classifiers under imbalance (Saito and Rehmsmeier, 2015).

6.3 Results of Data-Level Strategies

Data-level balancing strategies significantly improved minority class detection. By modifying the dataset distribution, classifiers were exposed to more spam samples during training, resulting in improved recall and F2-score.

6.3.1 RUS Results

Random Undersampling improved recall compared to baseline by reducing majority class dominance. However, because many legitimate emails were removed, some classifiers experienced reduced precision. While RUS enhanced spam detection sensitivity, its information loss sometimes negatively affected overall generalization. F2-score improved moderately but was not consistently stable across datasets (He and Garcia, 2009).

6.4 Results of Cost-Sensitive Strategies

Cost-sensitive strategies improved performance without modifying dataset distribution. These approaches primarily enhanced recall while maintaining better control over false positives compared to data-level methods.

7. CONCLUSION

This research presented a comparative evaluation of data-level balancing techniques and cost-sensitive learning strategies for spam email classification under highly imbalanced dataset conditions. Experiments were conducted on three benchmark datasets, namely SpamAssassin, Ling-Spam, and CSDMC2010, with simulated imbalance ratios of 90:10, 95:5, and 98:2 to reflect real-world spam filtering environments. The baseline classifiers showed strong performance degradation under extreme skewness, where accuracy remained high but recall and F2-score dropped significantly, confirming the presence of the accuracy paradox in imbalanced spam datasets. Data-level methods such as Random Undersampling and Random Oversampling improved minority class detection but exhibited limitations such as information loss and overfitting. SMOTE achieved the most consistent improvement among resampling techniques by generating synthetic spam samples, leading to higher recall and F2-score values. Cost-sensitive strategies, including class-weighted SVM, MetaCost, and threshold tuning, demonstrated stable performance without modifying dataset distribution. Among these, threshold optimization proved highly effective in improving F2-score by increasing spam recall while maintaining acceptable precision. The comparative results indicate that SMOTE-based resampling is effective when maximizing spam detection is the primary objective, whereas threshold tuning and class-weighted learning are more suitable when minimizing false positives is critical. Overall, this study provides a structured framework for selecting imbalance-handling techniques in spam filtering systems, highlighting the importance of using imbalance-aware metrics such as F2-score and AUC-PR for realistic evaluation.

8. FUTURE SCOPE

Future work can extend this research by incorporating deep learning-based architectures such as CNN-LSTM, Bi-LSTM, and Transformer models, which can capture contextual

relationships beyond TF-IDF features. Advanced oversampling techniques such as ADASYN and Borderline-SMOTE may also be explored to generate more realistic minority samples and reduce class overlap. Additionally, concept drift handling should be investigated since spam patterns evolve continuously over time, making static models less reliable in real-world deployment. Future studies may also develop adaptive cost-sensitive frameworks where misclassification costs dynamically change based on user feedback and email priority. Finally, hybrid models combining resampling with cost-sensitive learning could be designed to achieve better trade-offs between spam recall and false positive minimization in highly skewed environments.

REFERENCES

1. Androutsopoulos, I., Koutsias, J., Chandrinou, K.V. and Spyropoulos, C.D. (2000) 'An evaluation of naive Bayesian anti-spam filtering', in Proceedings of the Workshop on Machine Learning in the New Information Age. Barcelona, Spain, pp. 9–17.
2. Bergstra, J. and Bengio, Y. (2012) 'Random search for hyper-parameter optimization', *Journal of Machine Learning Research*, 13, pp. 281–305.
3. Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32.
4. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.
5. Domingos, P. (1999) 'MetaCost: A general method for making classifiers cost-sensitive', in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99). San Diego, CA: ACM, pp. 155–164.
6. Drummond, C. and Holte, R.C. (2006) 'Cost curves: An improved method for visualizing classifier performance', *Machine Learning*, 65(1), pp. 95–130.
7. Elkan, C. (2001) 'The foundations of cost-sensitive learning', in Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI). Seattle, WA, USA, pp. 973–978.
8. Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters*, 27(8), pp. 861–874.
9. Han, H., Wang, W.Y. and Mao, B.H. (2005) 'Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning', in Proceedings of the International Conference on Intelligent Computing (ICIC). Berlin: Springer, pp. 878–887.
10. He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284.
11. Japkowicz, N. and Shah, M. (2011) *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge: Cambridge University Press.
12. Joachims, T. (1998) 'Text categorization with support vector machines: Learning with many relevant features', in Proceedings of the 10th European Conference on Machine Learning (ECML). Berlin: Springer, pp. 137–142.
13. Manning, C.D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
14. Metsis, V., Androutsopoulos, I. and Paliouras, G. (2006) 'Spam filtering with Naive Bayes – Which Naive Bayes?', in Proceedings of the Third Conference on Email and Anti-Spam (CEAS 2006). Mountain View, California, USA.
15. Pedregosa, F. et al. (2011) 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.
16. Powers, D.M.W. (2011) 'Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation', *Journal of Machine Learning Technologies*, 2(1), pp. 37–63.
17. Saito, T. and Rehmsmeier, M. (2015) 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLOS ONE*, 10(3), e0118432.
18. Sculley, D., O'Connor, M., McCallum, A. and Corrada-Emmanuel, A. (2011) 'Spam filtering using machine learning techniques', in Proceedings of the Conference on Email and Anti-Spam (CEAS).
19. Veropoulos, K., Campbell, C. and Cristianini, N. (1999) 'Controlling the sensitivity of support vector machines', in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden, pp. 55–60.