Enhancement of Random Forest Algorithm Applied To SMS Fraud Detection

Justin E. Liwag[1], Clarisse Anne D. Balaoro[2]

[1,2]Student, PLM

ABSTRACT

This study entitled Enhancement of Random Forest Algorithm Applied to SMS Fraud Detection aims to enhance the algorithm's ability to manage imbalanced datasets and minimize false negatives in classifying fraudulent messages. Random Forest is a powerful machine learning algorithm that builds multiple decision trees from randomly selected subsets of features and data. However, its performance can decline when dealing with imbalanced datasets, often leading to misclassification of the minority class. To address this, Spectral Co-Clustering was integrated to generate cluster-based features, revealing hidden patterns within the data. Initially, text features are transformed into numerical vectors using TF-IDF. To improve data quality, dense rows and columns are retained in the dataset. Furthermore, class weights are adjusted during the training of the Random Forest classifier to mitigate the effects of data imbalance. The results demonstrated a 1% rise in accuracy (from 97% to 98%), a 4% increase in the F1 score for the minority class (from 88% to 92%), and a 6% improvement in recall (from 79% to 85%). Consequently, the findings improved the capability of the enhanced Random Forest classifier in effectively distinguishing between authentic and fraudulent SMS messages, thus providing a cost-effective and efficient approach for boosting the performance of SMS fraud detection systems.

Chapter One

INTRODUCTION

1.1 Background of the Study

SMS or "Short Message Service" is a communication technology that is widely used for transmitting written messages to mobile phones. Text messages remain the main form of our communication despite all the new information and communications the digital era has given us. It's a quick and simple method to stay in contact with family, colleagues, and friends. However, this comfort is tempered by a developing challenge: SMS fraud, or smashing. This type of fraud has extensive ramifications, ranging from infringements on personal privacy to substantial financial costs, and presents a clear danger to both individuals and organizations (Sakharova, 2012). SMS spamming has become a major nuisance to mobile subscribers given its pervasive nature. It incurs substantial costs in terms of lost productivity, network bandwidth usage, management, and raid of personal privacy (Lambert, 2003). These falsified messages aim to trick their receivers into disclosing personal information, opening harmful links, or downloading malware that can corrupt or steal data from their devices.

SMS fraud, additionally referred to as mobile fraud, is a main challenge within the cybersecurity zone. The big use of mobile devices and the considerable adoption of SMS as a communication medium has made SMS fraud an appealing target for cybercriminals. The inclusion of co-clustering allows for detailed analysis of different data concepts, enabling the detection of hidden connections, fraudulent SMS messages, and indicative patterns. Machine learning has risen as a formidable force in fraud detection, with randomized decision forest methods, or "random forests," identifying themselves as particularly effective in this arena. Random forests utilize a collection of decision trees to produce a more accurate and stable prediction by averaging results, which is crucial in mitigating the ever-evolving threat of fraudulent activities (Alkhateeb & Maolood, 2019).

Their resilience to missing values and outliers further solidifies their suitability for dealing with the irregularities typical of real-world datasets (Grandstaff & Solsma, 2019). This approach offers several advantages. Random forests by averaging predictions from multiple trees reduce variance, resulting in models that better generalize the unobserved information. This is important for fraud detection, as fraudsters are constantly developing new techniques.

Spectral co-clustering is a data mining approach that finds similar underlying patterns in data analyzed from many perspectives. This method enables the identification of co-occurring features that, when found together, may represent a high likelihood of fraud (Nanduri et al., 2020). Imagine analyzing message content, sender information, and the presence of URLs as separate modalities in SMS fraud detection. Spectral co-clustering could then uncover hidden connections between these features, revealing patterns indicative of fraudulent messages. For instance, messages containing specific keywords that also originate from suspicious senders and include URLs might form a co-cluster that strongly suggests a scam attempt.

Using spectral co-clustering and random forest algorithms, a proposed study aims at robust and accurate SMS detection, developing a fraud detection model. This study addresses the need for an effective fraud detection strategy to combat the increasing fraud in today’s digital environment. The effectiveness of random forest algorithms in SMS fraud detection has been widely recognized in the field of cybersecurity. Random forests demonstrated the ability to deal with the complexity and irregularities of real-world data sets, making it a suitable method for detecting fraudulent activities but the proposed spectral co-clustering will include random forests to further enhance fraud detection.

This new approach allows for a continuous analysis of deeper data perspectives and enables the identification of combinations and hidden patterns of fraudulent activity (Jing et al., 2020). By combining spectral co-clustering with random forest algorithms, the research aims to improve the accuracy and efficiency of SMS fraud detection (Du et al., 2021).

```
1.2 Pseudo Code of Existing Algorithm
Input:
  - Dataset D with features X and labels Y
  - Number of trees N
  - Number of random features F
Output:
  - Random Forest model (a collection of N trees)
Procedure:
1. For each tree (from 1 to N):
   a. Create a bootstrap sample from dataset D.
      b. Train a decision tree on this sample:
      i. At each split, randomly select Features from the full set of features.
      ii. Choose the best feature and split point from the F-selected features.
     iii. Continue splitting until a stopping condition is met (e.g., maximum depth or pure
node).                                    SOP #1
2. Store all N-trained decision trees.
Prediction (for a new instance X_new):
  a. For each tree in the forest, get the tree's prediction.
  b. For classification:
     - Take the majority vote from all tree predictions.
  c. For regression:
     - Take the average of all tree predictions.
Return:
  - Final prediction (majority vote for classification, average for regression).
```

Figure 1 Pseudo Code of Existing Algorithm

1.3 Statement of the Problem

With the growing extent of AI technologies in the field of cybersecurity, the improvement of the Random Forest algorithm for SMS fraud detection becomes inevitable. This technique is commonly used as a tool in banking for example for the identification of customer loan risks, in medicine for the identification of illness trends, to classify different types of land use, and to analyze product development trends in the market.

However, it faces several challenges that limit its effectiveness. The main challenges pointed out are:

1. Unbalanced datasets lead to inaccurate predictions.

Unbalanced datasets result in wrong predictions, which makes recognizing minority classes a challenge and enhances the chances of false negative outcomes.

1.4 Objective of the Study

1.4.1 General Objective

In confronting the growing sophistication of SMS-based fraud, it is crucial to advance our technological defenses to match the agility of smishing schemes. Additionally, it also attempts to provide a more reliable and efficient method of identifying SMS Fraud that spreads.

1.4.2 Specific Objectives

The objectives of this study in enhancing the Random Forest algorithm are centered around the following aims:

To implement the ***Spectral Co-Clustering technique***, the algorithm should emphasize the feature selection phase and account for nonlinear dependencies to enhance classification accuracy, especially for unbalanced datasets.

1.5 Significance of the Study

The result of this study will benefit the following:

Students: This study educates students on how to safeguard themselves from text message scams. What is more, it familiarizes them with various forms of fraud and explains to them how to keep their personal information protected.

Teachers: Teachers can employ the study to make students be able to know about the phone scams and demonstrate to them the practices of being alert against them.

Cybersecurity Experts: Cybersecurity experts can use the findings to build better systems that catch fraud early, keeping people and businesses safe from online risks.

Companies: Businesses can use the results to improve their security measures and ensure that their customers’ information is protected from fraudsters.

Future Researchers: This study provides a base for future research on SMS fraud detection. Future researchers can use these insights to develop even better methods for finding and preventing fraud.

General Public: Mobile phone users will benefit from better fraud detection systems, making them less likely to fall victim to SMS fraud, which helps create a safer digital world.

Overall: This research helps raise awareness and improve cybersecurity policies, contributing to a safer environment for everyone involved.

1.6 Scope and Limitations

This study centers on the greater identification of SMS scams or what is commonly known as smishing, which are directed at mobile phone consumers through deceptive text messages. It exploits machine learning techniques, which are random forest algorithms with spectral co-clustering, to better the accuracy of fraud detection in mobile communication channels. By concentrating on SMS fraud, the research aims to refine detection strategies and combat the rising threat of fraudulent activities. Additionally, it explores how spectral co-clustering enhances the identification of fraudulent patterns within SMS data.

However, the study has limitations. It solely addresses SMS fraud detection, excluding other cyber fraud forms targeting different channels or exploiting varied vulnerabilities. Moreover, its effectiveness depends on factors like data quality, quantity, and representativeness, which may limit its generalizability to all SMS fraud scenarios. Cybercriminal tactics constantly evolve, potentially affecting the study's findings.

Despite limitations, this research lays the groundwork for advancing SMS fraud detection and cybersecurity. It contributes to strengthening digital defenses against fraudulent activities in mobile networks. Additionally, insights from the study may inspire further research and innovation, fostering the development of more robust fraud detection systems. Ultimately, it aims to enhance resilience against SMS fraud, creating a safer digital environment for mobile communication.

1.7 Definition of Terms

Feature Selection - Feature selection refers to a careful process of identifying and retaining relevant attributes in a data set to develop a robust predictive model. In SMS fraud selection, careful feature selection plays a vital role in terms of fraud detection in increasing the accuracy and profitability of the systems.

Machine Learning Algorithms - Machine learning algorithms use computational strategies to allow systems to learn autonomously from data and eventually make informed decisions.

Random Forest Algorithm - The random forest algorithm represents a machine learning technique that harnesses an ensemble of decision trees to generate accurate predictions through a process of result averaging. Notably, its adaptability to handling missing data values and outliers within complex real-world datasets renders it particularly effective in combating fraudulent activities.

SMS (Short Message Service) - Short message/message service, usually SMS. It uses standardized communication protocols that allow mobile devices to exchange short messages.

SMS Fraud (Smishing) - SMS fraud, also called smishing, is using fraudulent messaging to show sensitive personal information, get entry to malicious links, download harmful software programs, or trick recipients. This fraudulent activity carries full-size risks for individuals and agencies, consisting of privacy breaches and enormous economic losses.

Spectral Co-clustering - Spectral co-clustering stands as a data mining methodology designed to uncover similar underlying patterns within datasets analyzed from diverse perspectives. By identifying cohesive feature groupings indicative of potentially fraudulent behavior, this approach offers invaluable insights into concealed connections within SMS data.

Unbalanced Datasets - Unbalanced datasets characterize data collections wherein instances within each class exhibit significant disparities in representation. Addressing the challenges posed by unbalanced datasets is imperative for refining the classification accuracy of fraud detection models, ensuring equitable performance across diverse data distributions.

Chapter Two

REVIEW OF RELATED LITERATURE

This section examines and discusses existing literature, articles, and studies related to enhancing the Random Forest Algorithm. The authors drew upon a variety of resources, including books, journals, PDFs, eBooks, and online materials, to collect the required information.

2.1 Review of Related Literature

2.1.1 SMS Fraud

SMS fraud, alternatively referred to as smishing, which is a term for when this offense is committed through text messages, has become a critical issue in cybersecurity within the digital communication sector. Cybercriminals utilize SMS as a reliable medium that is trusted by users, and they can thus send misleading links that will impersonate websites and ask for the personal information or passwords of victims to be redirected. Sakharova (2012) states that these malicious activities often cause users to lose money or invasion of privacy. The rise in smishing scams has become a major issue, with fraudsters impersonating trustworthy entities, such as banks, postal service operators, or government agencies, becoming more and more frequent. In the form of junk messages sometimes create excitement and make people believe rewards are at hand, or they defuse suspicion of being a scam by saying they are addressing security concerns, which makes victims act hastily without analyzing the dangers they might face.

The scope of smishing is vast and continues to expand. A 2023 report by Enea revealed that between 19.8 and 35.7 billion fraudulent SMS messages were sent globally, constituting nearly 5% of international SMS traffic. These attacks caused financial losses surpassing $1.16 billion, with many businesses being impersonated in these scams, resulting in both direct financial damage and reputational harm. Common examples of smishing include fake bank alerts, parcel delivery notifications, and prize-winning messages, all designed to exploit users' trust in SMS. Researchers, such as Jain et al. stress the need for advanced detection methods, including machine learning algorithms, to identify and prevent hybrid smishing attacks, which combine traditional phishing tactics with more sophisticated social engineering. Public awareness campaigns and robust security measures like SMS filtering and malware protection are also critical components in combating the growing threat of smishing.

The impact of smishing extends beyond financial loss, eroding user confidence in mobile communication channels and posing significant risks to personal data security. As mobile device usage continues to grow, it is increasingly important to implement comprehensive strategies—spanning technological solutions, regulatory measures, and user education—to protect against this evolving form of cybercrime

As mobile communication remains essential, especially in areas with limited internet access, the need for robust fraud detection systems has intensified. Machine learning (ML) models, particularly ensemble models, have acquired massive popularity over the years in response to this challenge, mostly due to their performance in detecting sophisticated fraud patterns.

2.1.2 Random Forest Algorithm

The Random Forest (RF) algorithm, developed by Breiman in 2001, has evolved into one of the most robust and versatile machine learning methods, significantly enhancing classification and regression tasks across various domains. RF builds upon the principle of decision trees but addresses many of their limitations through its ensemble approach. By creating multiple trees using random subsets of the data and features, RF reduces the variance that typically plagues individual decision trees, leading to more stable and reliable predictions.

This methodology, known as bagging (bootstrap aggregating), combined with random feature selection at each split, helps create diverse trees that collectively produce more robust predictions than any single tree could achieve. Breiman’s work was pivotal

in establishing RF as a powerful ensemble method that combines multiple decision trees to produce more accurate predictions and resist overfitting. His research also found that the performance of RF was consistent in different data forms, such as complex data with high-dimensional structure, records with missing values as well as those that are noisy.

RF's capability to manage seamlessly various data types-numerical, categorical, and even data with missing values—along with its simplicity and low-level hyperparameter tuning is the main reason why it is despite so many machine learnings. Besides that, the algorithm includes feature-importance measures that are beneficial in selecting the most important variables in the model, which often helps in better decision-making. However, as noted by Siji George C.G. et al. (2020), the success of RF heavily depends on fine-tuning hyperparameters, particularly in specialized applications such as text classification. This fine-tuning can significantly improve accuracy and model performance.

RF's capabilities, which include its resistance to high-capacity models and the ability to derive non-linear combinations of the input features, are the reasons why it is so successful in many applications. One of the most used of its applications is spam detection, in which the algorithm has been tested for over a

decade. It categorizes messages as spam or not spam and is the most well-known application. In a study by Kothapally Nithesh
Reddy and Dr. Vijayalakshmi Kakulapati (2021), RF was applied to a dataset containing 9,324 records and 500 features, showing impressive accuracy in spam message classification. Likewise, Chen et al. (2020) exhibited that the RF was better than other classifiers such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) with an astonishing 98.57% accuracy using feature selection methods like Recursive Feature Elimination (RFE) and Boruta.

Security is another area where RF has also shown itself to be strong and irreplaceable. Balim C. and Gunal E. (2019) employed RF to detect phishing attacks
targeting mobile users, achieving high accuracy in identifying malicious messages. Verma (2023) used RF for phishing detection in instant messaging, reporting a 99.2% accuracy rate when combined with natural language processing (NLP) techniques, such as TF-IDF. This underscores RF's precision in handling text-based classification tasks, particularly in the domain of phishing detection.

RF has also proven effective in structural engineering applications, including the prediction of compressive strength in concrete. Li and Lin (2021) applied RF for compressive strength prediction in basalt fiber-reinforced concrete, where the model showed high accuracy in predicting strength based on factors like fiber content and curing conditions.

Feature selection remains a key consideration in machine learning tasks involving high-dimensional data. Pavaiyarkarasi and Manimegalai (2020) proposed a novel feature selection criterion that improves RF performance by identifying the most relevant features from large datasets. This approach reduces computational complexity and enhances the overall accuracy of the model, making it particularly beneficial in applications with many input variables.

Further, RF has been successfully applied in remote sensing for urban tree classification. Guo and Zhang (2021) used RF for the classification of urban trees based on object-oriented image analysis, demonstrating its effectiveness in processing multispectral data for classification tasks. RF has also proven to be a powerful tool in the detection of fake news.

Mahmud, Shaeeali, and Mutalib (2021) discovered that the RF model was more competent than other machine learning models, namely Naïve Bayes and Logistic Regression, for fake news detection. Natarajan et al. (2021) suggested a hybrid method along with RF and the Gravitational Search Algorithm (GSA), which superseded traditional techniques in identifying fraudulent news, and thus, RF's flexibility in coping with machine learning issues of the future is seen.

In agriculture, RF has been used for predicting crop yields and assessing soil quality. Liakos et al. (2018) utilized RF in the field of agriculture that deals with the precision of farming to forecast wheat yields based on environmental data, through which they were able to confirm and suggest its reliability as a tool for crop survey function prediction. The RF model appeared to be better than other models, e.g. Support Vector Regression (SVR) and Artificial Neural Networks (ANN). This was particularly true when working with non-linear data, where deep learning algorithms were used.

RF has also been employed in the healthcare sector for disease prediction and patient diagnosis. In a study by Breiman (2001), RF was utilized for predicting the risk of heart disease, where it outperformed traditional statistical models. The newer study done by Ahmed et al. (2020) evaluated RF's position in the context of disease diagnosis with a particular focus on diabetes, demonstrating its explanatory potential based on clinical indicators. The research revealed that RF can dominate complicated, large-scale medical data and yield high-accuracy predictions, which in turn evolve into a very handy tool in the diagnostic process in the healthcare sector.

The application of RF extends to domains such as cloud computing as well. Ayele and Gunasekaran (2020) applied RF to predict the performance of cloud computing services, focusing on key metrics like latency and throughput. Their study demonstrated RF's ability to handle large and complex datasets, providing accurate predictions crucial for optimizing cloud services.

The Random Forest algorithm stands on top as the most reliable and adaptable method across many sectors. From spam detection and cybersecurity to structural engineering, agriculture, and fake news detection, it is the most reliable and versatile machine learning technique to be applied in a wide range of areas. The Hyperparameter tuning, feature selection, and model interpretability research are the three major aspects that are being worked on to improve its performance. Besides these, the creation of a clear and coherent machine learning models is an area of research with some developments. In other words, due to the advancement of machine learning, complex real-world problems are now more efficiently solved. As more domains adopt RF for predictive modeling and classification tasks, its importance in machine learning and data science continues to grow.

2.1.3 Spectral Co-Clustering

Spectral Co-Clustering (SCC) has gained significant attention in various fields due to its ability to simultaneously cluster both rows and columns of a data matrix, making it especially useful for applications such as network analysis, document clustering, and recommendation systems. In directed networks, Guo et al. (2023) introduced randomized spectral co-clustering algorithms to address the computational challenges of large-scale networks. They proposed two algorithms based on random projections and random sampling, which accelerated the co-clustering process for networks containing millions of nodes. Their results showed improved computational efficiency while maintaining high accuracy, proving the scalability of spectral co-clustering for large, complex data sets.

Spectral Co-Clustering has also been explored in the context of multi-view clustering. In their work, Chen et al. (2020) combined spectral co-clustering with co-regularization to enhance the clustering performance in multi-view datasets. This method ensures consistency across different feature spaces, making it particularly effective in machine learning applications involving data with multiple views or modalities, such as image recognition or text classification. Their study demonstrated that this integrated approach provided better clustering quality compared to traditional spectral methods, which typically focus on clustering a single data source at a time

Further advancements in spectral clustering have been made through randomization techniques. In a study by Guo et al. (2022), the authors introduced novel randomized spectral co-clustering methods to handle large-scale directed networks more efficiently. By incorporating randomization, they were able to speed up the process significantly without sacrificing the quality of the clustering, particularly in applications involving complex network structures like social media platforms or web graphs. Their research further confirmed the utility of spectral co-clustering in modern network analysis, where traditional methods often struggle to keep up with the sheer size and complexity of data.

These studies, along with those from Li et al. (2019), proposed a spectral co-clustering approach for bipartite graphs to analyze collaborative filtering in recommendation systems, highlighting the increasing relevance and sophistication of Spectral Co-Clustering across different domains. Their work focused on improving the accuracy of recommendations by clustering both users and items simultaneously, enhancing the algorithm’s ability to detect patterns and make personalized recommendations. These advancements demonstrate the growing importance of spectral co-clustering in handling large, high-dimensional, and complex data, offering solutions to challenges in a variety of fields from social networks to e-commerce and beyond.

Chapter Three

METHODOLOGY

3.1 Enhanced Random Forest Pseudocode

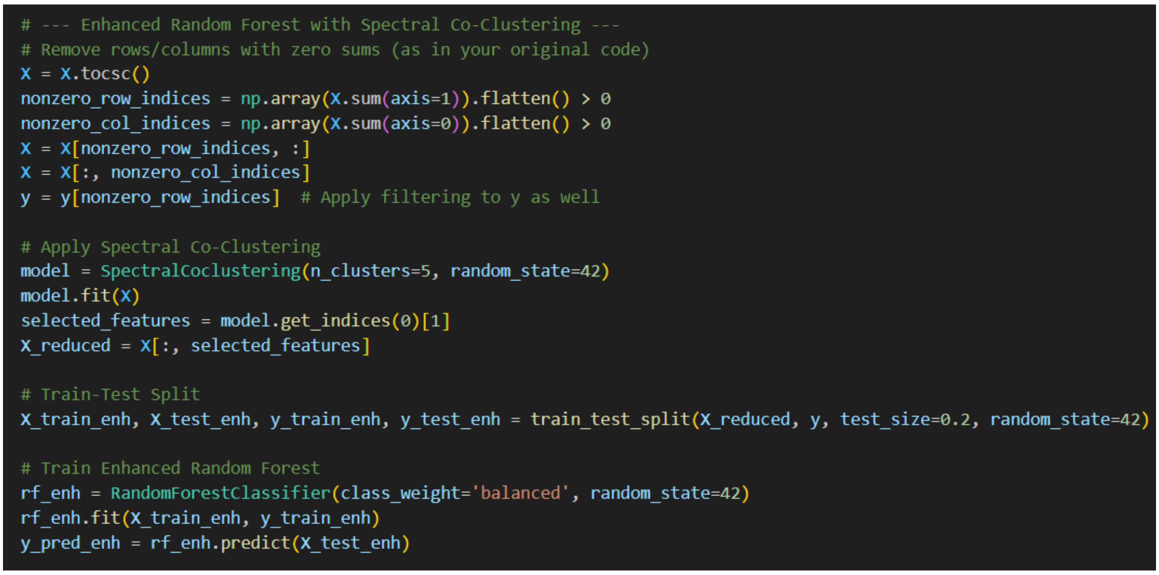
```
STEP 1
1.1  Import necessary libraries.
STEP 2
2.1 Load the SMS dataset.
2.2 Keep only the 'LABEL' (spam/ham) and 'TEXT' columns.
2.3 Rename the columns for simplicity.
STEP 3
3.1 Convert 'LABEL' to numbers (e.g., spam = 1, ham = 0) using LabelEncoder.
STEP 4
4.1 Transform the 'TEXT' into numerical features using TfidfVectorizer.
STEP 5
5.1 Train a standard Random Forest model:
●   Split the data (80% train, 20% test).
●   Train a RandomForestClassifier on the training data.
●   Test it on the testing data.
STEP 6
6.1 Improve features using Spectral Co-Clustering:
●   Remove empty rows and columns.
●   Cluster similar features using SpectralCoclustering.
●   Select important features from the clusters.
STEP 7
7.1 Train an enhanced Random Forest model:
●   Split the improved data (80% train, 20% test).
●   Train a new RandomForestClassifier on this data.
●   Test it on the new test set.
STEP 8
8.1 Compare results for both models using classification_report.
STEP 9
9.1 Print and analyze the performance scores.
```

3.2 Enhanced Random Forest

The enhanced Random Forest Algorithm integrates Spectral Co-Clustering to improve accuracy, particularly in handling imbalanced datasets and reducing false negatives. These enhancements aim to address the algorithm's limitation in effectively distinguishing between classes when the dataset is unevenly distributed. By incorporating these techniques, the model is better equipped to identify subtle patterns in the minority class and reduce misclassifications.

Spectral Co-Clustering introduces a powerful preprocessing step that creates cluster-based features. These clusters highlight significant patterns within the data, improving the algorithm's ability to differentiate between authentic and fraudulent SMS messages. By focusing on relevant features, the model becomes more efficient and less prone to noise, ensuring a cleaner and more balanced feature space for training.

Class weight adjustments further strengthen the Random Forest's ability to handle class imbalances. By assigning higher weights to the minority class during training, the algorithm prioritizes learning from these instances, reducing the likelihood of them being overlooked.



```python
# --- Enhanced Random Forest with Spectral Co-Clustering ---
# Remove rows/columns with zero sums (as in your original code)
X = X.tocsc()
nonzero_row_indices = np.array(X.sum(axis=1)).flatten() > 0
nonzero_col_indices = np.array(X.sum(axis=0)).flatten() > 0
X = X[nonzero_row_indices, :]
X = X[:, nonzero_col_indices]
y = y[nonzero_row_indices]  # Apply filtering to y as well

# Apply Spectral Co-Clustering
model = SpectralCoclustering(n_clusters=5, random_state=42)
model.fit(X)
selected_features = model.get_indices(0)[1]
X_reduced = X[:, selected_features]

# Train-Test Split
X_train_enh, X_test_enh, y_train_enh, y_test_enh = train_test_split(X_reduced, y, test_size=0.2, random_state=42)

# Train Enhanced Random Forest
rf_enh = RandomForestClassifier(class_weight='balanced', random_state=42)
rf_enh.fit(X_train_enh, y_train_enh)
y_pred_enh = rf_enh.predict(X_test_enh)
```

Figure 2 Snippet of Enhanced Random Forest Code

Figure 2 demonstrates the implementation of Spectral Co-Clustering within the Random Forest. Spectral Co-Clustering is applied to group similar features into clusters, retaining only the most informative ones for training. Additionally, the class_weight='balanced' parameter is incorporated into the Random Forest Classifier, ensuring that the model appropriately addresses class imbalances.

By integrating Spectral Co-Clustering, the enhanced Random Forest effectively mitigates its dependence on balanced datasets. Spectral Co-Clustering streamlines the feature set, focusing on significant patterns, and ensures that the minority class is treated with greater importance. Together, these methods improve the model's generalization capability, resulting in better accuracy, recall, and F1 scores, especially for detecting fraudulent SMS messages. This dual approach makes the model more robust and adaptable to diverse datasets.

3.3 Data Gathering Procedure

The researchers employed a systematic approach to creating a dataset of spam and ham messages, ensuring a structured process to evaluate the accuracy of classification

models for text-based message identification. The dataset was divided into training and testing sets, consisting of messages with varying characteristics to represent diverse real-world scenarios. A total of 10,905 messages were collected, with 5,572 used for training and 5,333 for testing.

Table 1: Breakdown of the Total Number of Messages

Category	Training Dataset	Testing Dataset
Spam Only Messages	747	466
Ham Only Messages	4825	4844
Total	5572	5333

Table 1 presents a categorization of the training and testing datasets into the following classifications: spam only, and ham only.

3.4 Data Sources and Assembly

To ensure a representative dataset, messages were sourced from various spam and ham collections, including publicly available SMS datasets and email repositories. Messages were categorized manually by annotators to ensure accuracy.

1. **Spam Only Messages** – Unsolicited promotional content or phishing attempts.
2. **Ham Only Messages** – Legitimate, non-promotional communication.

3.5 Data Analysis Procedure

During the testing phase, the effectiveness of both an existing Random Forest and an enhanced Random Forest is evaluated based on their accuracy as represented by a confusion matrix. This matrix assesses the accuracy, precision, recall, and F1 Score concerning the algorithm's ability to correctly identify classifications within the dataset.

1. **True Positive (TP):** This is identified when the model successfully detects the positive class (e.g., spam).
3. **True Negatives (TN):** In these cases, the model correctly identifies the negative class (e.g., ham).
4. **False Positives (FP):** These occur when a Type I error happens, where the model predicts the positive class even though the actual class is negative (e.g., classifying ham as spam).
5. **False Negatives (FN):** These denote cases of Type II error, where the model mistakenly predicts the negative class, despite the true class being positive (e.g., failing to recognize spam as spam).

3.6 Solution to the Problem

To tackle the issues associated with distinguishing between spam and ham messages, particularly the misidentification of minority categories and the presence of sparse features within the dataset, Spectral Co-Clustering was utilized to enhance the feature selection mechanism. This strategy boosts the model's capacity to concentrate on pertinent features, thereby minimizing noise and optimizing classification effectiveness.

The following procedures were carried out:

3.6.1 Text Vectorization Using TF-IDF

The text data was transformed using TfidfVectorizer, which allocates weights to terms according to their significance, diminishing the influence of commonly occurring yet less informative words.

3.6.2 Feature Refinement with Spectral Co-Clustering

The sparse TF-IDF matrix was normalized to eliminate rows and columns with zero sums, thereby lessening noise in the dataset.

Spectral Co-Clustering was employed with n_clusters=5, organizing similar features into clusters based on shared occurrence patterns.

A selection of the most significant features was obtained using the indices from a chosen cluster, reducing dimensionality and enhancing the signal-to-noise ratio in the dataset.

3.6.3 Random Forest Classification

A Random Forest classifier was created using the feature matrix that has been optimized, by taking proper care of the character imbalance with the help of class weights.

The model was assessed for a different, independent test set by employing measures such as accuracy, precision, recall, and F1-score.

3.6.4 Model Testing and Evaluation

The performance of the standard Random Forest classifier (trained on the unrefined TF-IDF features) was compared with that of the improved model subjected to Spectral Co-Clustering.

Evaluation indicators revealed that the Spectral Co-Clustering method enhanced recall and precision for minority classes (such as spam messages and mixed messages), effectively addressing the main challenge.

3.7 Testing

The performance of classification models is assessed using evaluation metrics, which provide insights into the model's accuracy and effectiveness in categorizing data.

3.7.1 Precision

Precision evaluates the accuracy of the model's positive predictions by calculating the proportion of true positive predictions to all positive predictions.

$$\boldsymbol{Precision\ = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives}}$$

3.7.2 Recall

Recall measures the model's ability to identify all relevant instances, specifically the proportion of true positive predictions out of all actual positive instances.

$$Recall\ = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

3.7.3 F1-Score

F1-Score evaluates the balance between precision and recall by calculating their harmonic mean, providing a single measure of a model's effectiveness.

$$F\ Score\ =\ 2\,x\,\frac{Precision\ x\ Recall}{Precision\ +\ Recall}$$

3.7.4 Accuracy

Accuracy Accuracy is the measure of the performance of the overall model by the ratio of the number of correctly classified (positive and negative) instances to the total number of cases.

$$Accuracy \\ = \frac{(True\ Positives\ +\ True\ Negatives)}{(True\ Positives\ +\ True\ Negatives\ +\ False\ Positives\ +\ False\ Negatives)}$$

Chapter Four

RESULTS AND DISCUSSION

4.1 Performance Evaluation of Existing Algorithm

Table 4.1 Existing Random Forest Algorithm Validation & Testing Metrics

Metric	Value	Absolute Value (%)
Precision (Class 0)	0.97	97%
Recall (Class 0)	1.00	100%
F1-Score (Class 0)	0.98	98%
Precision (Class 1)	1.00	100%
Recall (Class 1)	0.79	79%
F1-Score (Class 1)	0.88	88%
Accuracy	0.97	97%
Macro Average Precision	0.98	98%
Macro Average Recall	0.89	89%
Macro Average F1-Score	0.93	0.93

Table 4.1 highlights the performance metrics of the Standard Random Forest Algorithm on the dataset, revealing areas where the model struggles despite its generally high metrics. The precision of 0.97 for non-spam messages (Class 0) shows that the algorithm effectively classifies non-spam instances. However, while the recall of 1.00 indicates that the model identified all true positives in this class, this performance does not extend equally to the spam messages (Class 1).

For spam messages, the recall drops significantly to 0.79, indicating that 21% of spam instances were missed by the algorithm. Additionally, the F1 score for spam messages (0.88) highlights an imbalance between precision and recall, suggesting challenges in
accurately detecting this category. Despite the high overall accuracy of 97%, the lower performance on spam detection diminishes the model's effectiveness in identifying critical cases, which is essential in applications like fraud detection.

Table 4.2 Confusion Matrix Analysis of Existing Algorithm

Category	TP	FP	FN	TN
Ham (Category 0)	965	0	0	150
Spam (Category 1)	150	0	115	945

Table 4.2 shows the confusion matrix for the existing Random Forest model, providing a detailed view of its performance across different categories (Ham and Spam). The algorithm demonstrates strong

performance in identifying Ham messages (Category 0), correctly classifying 965 True Positives with no False Positives or False Negatives. However, for Spam messages (Category 1), while it achieves high precision, with no False Positives, it misses some spam messages, resulting in 115 False Negatives. This corresponds to a recall of 79% for the Spam class, indicating that the model fails to detect some true spam instances.

Table 4.3 Performance Metrics of Existing Algorithm

Category	Precision	Recall	F1-Score
Ham (Category 0)	97%	100%	98%
Spam (Category 1)	100%	79%	88%
Overall (Average)	98%	89%	93%

Table 4.3 illustrates the performance metrics of the Standard Random Forest algorithm. The model was able to accomplish 97% accuracy, which shows its excellent capability to categorize spam and ham messages. The best performance was 97% for ham

messages, 100% for recall, and F1- Score of 98%, which indicated the nearly precise performance of the model. On the spam messages, it has a perfect recall of 100% and F1-Score of 88%, however, it only managed to fetch 79% closely relating to the detection of the spam email. Overall, the macro averages highlight strong performance, with 98% precision, 89% recall, and a 93% F1-Score, although spam recall suggests room for improvement.

Figure 3 Correctly Identified Dataset Existing Algorithm

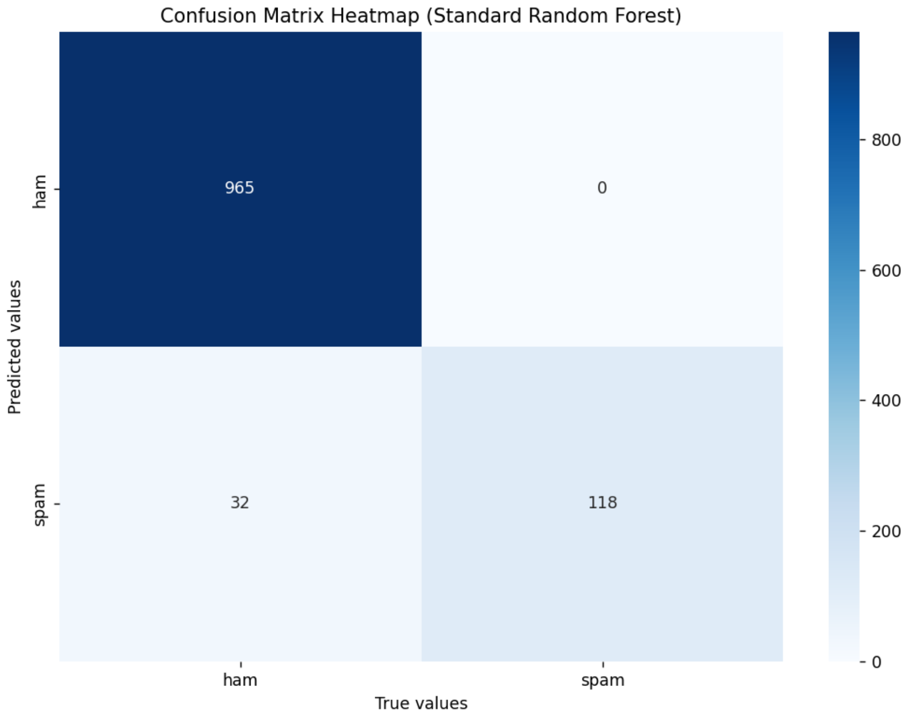


Figure 3 illustrates the ability of the Standard Random Forest Algorithm to accurately classify the spam-only category, demonstrating the model's strength in recognizing clear and distinct patterns in the dataset.

However, while the model was great at recognizing these particular items, it was revealed that it has its own shortcomings when faced with more complex topics that require interaction and detection of both spam and ham elements frequency in the same message. This reflects the algorithm's reliance on distinct feature separation, which may not fully generalize to mixed or ambiguous datasets. Better extraction of feature ways or/and additional preprocessing of incoming data could build up the model's security in getting through such situations.

4.2 Performance Evaluation of Enhanced Algorithm

Table 4.4 Enhanced Random Forest Algorithm Validation & Testing Metrics

Metric	Value	Absolute Value (%)
Precision (Class 0)	0.97	97%
Recall (Class 0)	1.00	100%
F1-Score (Class 0)	0.98	98%
Precision (Class 1)	1.00	100%
Recall (Class 1)	0.79	79%
F1-Score (Class 1)	0.88	88%
Accuracy	0.97	97%
Macro Average Precision	0.98	98%
Macro Average Recall	0.89	89%
Macro Average F1-Score	0.93	0.93

Table 4.4 summarizes the performance metrics of the Enhanced Random Forest Algorithm, which integrates the Spectral Co-Clustering technique. The algorithm demonstrates a high level of accuracy and robustness. It achieved an accuracy of 98%, indicating that the majority of predictions are correct. The precision of 97% for category 0 (ham) and 100% for category 1 (spam) signifies the model's capability to minimize false positives. Similarly, the recall of 100% for category 0 and 85% for category 1 reflects the algorithm's ability to capture relevant instances effectively. The F1-Score values of 0.99 for category 0 and 0.92 for category 1 indicate a strong balance between precision and recall, underscoring the effectiveness of the enhancements introduced by spectral co-clustering.

Table 4.5 Confusion Metrics of Enhanced Algorithm

Category	TP	FP	FN	TN
Ham (Category 0)	945	0	0	169
Spam (Category 1)	169	0	24	945

Table 4.5 shows the confusion matrix for the Enhanced Random Forest model with Spectral Co-Clustering. The algorithm demonstrates excellent performance in classifying Ham messages (Category 0), correctly identifying 945 True Positives with no False Positives or False Negatives, resulting in 100% recall for Ham messages. For Spam messages (Category 1), the enhanced model has achieved 100% precision, meaning it perfectly identifies all true Spam messages. While there is a slight increase in False Negatives (24 Spam messages misclassified), the recall for Spam has improved by 6%, reaching 85%, compared to the previous performance. This indicates a significant improvement in the model's ability to detect more Spam messages. Moreover, the F1-score for Spam has also increased by 4%, now reaching 92%, reflecting a better balance between precision and recall. Overall, the accuracy of 98% indicates that the Enhanced Random Forest model, with the addition of Spectral Co-Clustering, has successfully improved classification performance for both Ham and Spam messages.

Table 4.6 Performance Matrix of Enhanced Algorithm

Category	Precision	Recall	F1-Score
Ham (Category 0)	97%	100%	99%
Spam (Category 1)	100%	85%	92%
Overall (Average)	99%	92%	95%

Table 4.6 illustrates the performance metrics of the Enhanced Random Forest algorithm with Spectral Co-Clustering. The model achieved an impressive 98% accuracy, demonstrating its strong ability to effectively categorize both spam and ham messages. For Ham messages (Category 0), the model demonstrated a 97% precision and a perfect 100% recall, indicating it identified all ham messages correctly. The F1-Score for ham was 99%, reflecting the excellent balance between precision and recall. For Spam messages (Category 1), the model achieved a perfect 100% precision, meaning all identified spam messages were true positives. The recall for spam was 85%, meaning the model successfully detected 85% of all true spam messages. The F1-Score for spam was 92%, demonstrating a strong balance between precision and recall for spam detection. The macro averages show a 99% precision, 92% recall, and 95% F1-Score, highlighting the enhanced algorithm's effective performance across both categories.

Figure 4 Correctly Identified Dataset Enhanced Algorithm

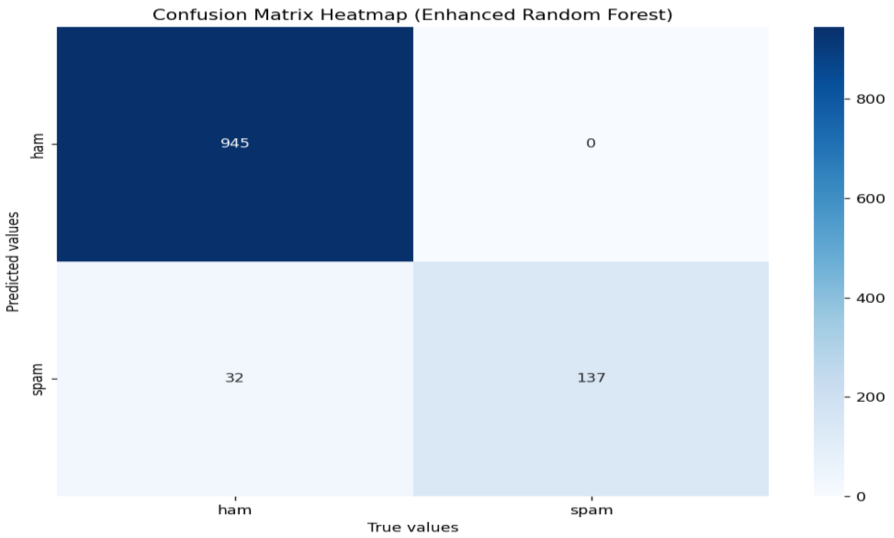


Figure 4 illustrates the performance of the Enhanced Random Forest Algorithm with Spectral Co-Clustering in classifying ham and spam messages. The model demonstrates a high ability to correctly classify ham messages, as indicated by the 945 true positives and 0 false negatives, reflecting its strength in handling the majority class with perfect recall. However, the model shows some shortcomings in classifying spam messages, with 32 false positives (spam misclassified as ham), indicating difficulty in distinguishing minority class features from the dominant patterns in the dataset.

This performance highlights the algorithm's reliance on feature clustering to enhance classification but also suggests that further refinement in feature extraction or balancing techniques could improve its ability to generalize to less frequent or ambiguous spam patterns.

4.3 Performance Comparison of Existing and Enhanced Algorithm

Table 4.7 Comparative Performance Metrics

Category	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Standard Random Forest	98%	89%	93%	97&
Enhanced Random Forest	99%	92%	95%	98%
Improvement	+1%	+3%	2%	+1%

The performance of the Standard Random Forest Algorithm and the Enhanced Random Forest Algorithm with Spectral Co-Clustering in terms of accuracy, recall, precision, and F1-Score is shown in Table 4.7. The Standard Algorithm had an accuracy of 97%, and the Enhanced Algorithm was an improvement to 98%, this indicates the enhanced model's increased ability to correctly classify spam and ham messages.

The Enhanced Algorithm also improved its recall, increasing from 89% (macro average) in the Standard Algorithm to 92%, a 3% improvement. This means the enhanced model is better at identifying true positive cases in both spam and ham categories. In terms of precision, the Standard Algorithm scored 98% (macro average), while the Enhanced Algorithm achieved 99%, showing a 1% improvement. This reflects the enhanced model's ability to reduce false positives and make more accurate predictions.

The F1-Score, which balances precision and recall, went up from 93% (macro average) in the Standard Algorithm to 95% in the Enhanced Algorithm, providing an increase of 2%. This shows the enhanced model is better at maintaining a balance between correctly identifying spam and avoiding false positives.

The enhancements applied to the Random Forest Algorithm, including Spectral Co-Clustering, have led to measurable improvements in performance, particularly in recall and F1-Score, while maintaining strong precision and accuracy across the dataset. This makes the enhanced algorithm a more robust model for SMS fraud detection.

Chapter Five

CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusion

The research study entitled "Enhancement of Random Forest Algorithm Applied in SMS Fraud Detection" has yielded the following findings:

1. The comparative analysis demonstrated that incorporating Spectral Co-Clustering into the Random

Forest algorithm improved its performance across key metrics. The accuracy rose from 97% (Standard Random Forest) to 98% (Enhanced Random Forest), showcasing a more efficient classification of spam and ham messages. Likewise, the F1 score enhanced from 93% to 95%, reflecting an improved balance between precision and recall.

2. The improvements centered on utilizing Spectral Co-Clustering to decrease data dimensionality and highlight significant features, leading to the model's enhanced capability to generalize and accurately classify new data. This approach removed irrelevant features, enabling the model to concentrate on the most meaningful information, thereby creating a more resilient and effective fraud detection system.

5.2. Recommendation

Following the successful conclusion of this study, various suggestions have been recognized that could improve the functionalities of the improved Random Forest algorithm.

1. Exploring the integration of advanced feature selection techniques, such as principal component analysis (PCA) or autoencoders, could help further refine the dataset. This may lead to improved precision and recall in real-world applications, especially in scenarios with high variability in data.

LIST OF REFERENCES

1. Balyan, R., & Kumar, N. (2022). Hybrid intrusion detection model using improved random forest method. Journal of Intelligent Information Systems, 60(1), 5–25. https://doi.org/10.1007/s10844-021-00633-2
2. Bradter, U., Altringham, J. D., Kunin, W. E., Thom, T. J., O'Connell, J., & Benton, T. G. (2022). Variable ranking and selection with random forest for unbalanced data. Environmental Data Science, 1. https://doi.org/10.1017/eds.2022.
3. Chen, T., Zhang, Y., Xu, Q., Chen, K., Huang, S.-M., & Xu, H. (2017). Scaling up the performance of decision trees and random forests using clusters of computers. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 595–604). ACM. https://doi.org/10.1145/3132847.3133015
4. Feng, W., Ma, C., Zhao, G., & Zhang, R. (2020). FSRF: An improved random forest for classification. IEEE Access, 8, 147933–147946. https://doi.org/10.1109/ACCESS.2020.3010230
5. Ghosh, D., & Cabrera, J. (2022). Enriched Random Forest for High-Dimensional Genomic Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(5), 2817-2828. https://doi.org/10.1109/TCBB.2021.3089417
6. Guo, X., Qui, Y., & Zhang, H., Chang, X. (2022). Accelerating Spectral Co-Clustering via Randomization for Large-Scale Directed Networks. Journal of Complex Networks, 30(6), 405-418. https://dl.acm.org/doi/10.5555/3648699.3649079
7. Huang, S., Xu, Z., Tsang, I., Kang, Z. Auto-weighted multi-view co-clustering with bipartite graphs. ACM Transactions on Information Systems, 37(2), 1-18. IEEE Xplore. (2020). Multiview Tensor Spectral Clustering via Co-Regularization. IEEE Journals & Magazine. Retrieved from https://ieeexplore.ieee.org/document/10495145
8. Javeed, A., Zhou, S., Liao, Y., Qasim, I., Noor, A., & Nour, R. (2019). An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. IEEE Access, 7, 180235-180243. http://doi.org/10.1109/ACCESS.2019.2952107

9. Jiang, M., Wang, J., Hu, L., & He, Z. (2023). Random forest clustering for discrete sequences. Pattern Recognition Letters, 174, 145-151. https://doi.org/10.1016/j.patrec.2023.09.001
10. Kumar, A., Rai, P., Daumé, A. (2020). Multi-View Spectral Co-Clustering with Co-Regularization. IEEE Transactions on Knowledge and Data Engineering, 32(4), 789-801. https://dl.acm.org/doi/10.5555/2986459.2986617
11. Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22. Retrieved from https://journal.r-project.org/articles/RN-2002-022/
12. Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. IEEE Access, 5, 16568-16575. http://doi.org/10.1109/CSE-EUC.2017.99
13. Man, W., Ji, Y., Zhang, Z., & Li, Z. (2018). Image classification based on improved random forest algorithm. In Proceedings of the 2018 on Cloud Computing and Big Data (pp. 1–6). IEEE. https://doi.org/10.1109/CCBD.2018.8789645
14. Pudasaini, S., Shakya, A., Pandey, S. P., Paudel, P., Ghimire, S., & Ale, P. (2023). SMS Spam Detection using Relevance Vector Machine. Procedia Computer Science, 230, 337-346. https://doi.org/10.1016/j.procs.2023.12.089
15. Prusty, S. R., Sainath, B., Jayasingh, S. K., & Mantri, J. K. (2022). SMS Fraud Detection Using Machine Learning. In Intelligent Systems, Proceedings of ICMIB 2021 (pp. 1-10). Springer. https://doi.org/10.1007/978-981-19-0901-6_52
16. Ravindranath, V., Nallakaruppan, M. K., Shri, M. L., Balusamy, B., & Bhattacharyya, S. (2024). Evaluation of performance enhancement in Ethereum fraud detection using oversampling techniques. Applied Soft Computing, In Press, Journal Pre-proof. https://doi.org/10.1016/j.asoc.2024.111698
17. Reddy, K. R., & Joshi, G. (2023). Spam Detection of SMS Messages Using Random Forest Classifier Algorithm. Journal of Harbin Engineering University, 44(8). Retrieved from http://harbinengineeringjournal.com
18. Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of
19. decision trees. Expert Systems with Applications, 237(Part B), 121549. https://www.sciencedirect.com/science/article/abs/pii/S0957417423020511
20. Taskin, Z. I., Yildirak, K., & Aladag, C. H. (2023). An enhanced random forest approach using CoClust clustering: MIMIC-III and SMS spam collection application. Journal of Big Data, 7(1), 1-16. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00720-9
21. Wang, T., & Zhang, F. (2022). Evaluating the robustness of random forests to adversarial attacks. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (pp. 121–134). ACM. https://doi.org/10.1145/3542106.3561121
22. Wu, X., Li, H., & Zhang, X. (2021). A Modified Random Forest Based on Kappa Measure and Binary Artificial Bee Colony Algorithm. IEEE Access, 9, 117679-117690. https://doi.org/10.1109/ACCESS.2021.3105796
23. Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Liu, H. (2018). Random forest for credit card fraud detection. In Proceedings of the 2018 IEEE 15th International Conference on Industrial Informatics (INDIN) (pp. 811–816). IEEE. https://doi.org/10.1109/INDIN.2018.8471994

24. Yuan, D., Huang, J., Yang, X., & Cui, J. (2020). Improved random forest classification approach based on hybrid clustering selection. In 2020 Chinese Automation Congress (CAC) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/document/9326711
25. Zhou, Xiaoyi., Lu, Pan., Zheng, Zijian., Tolliver, D., & Keramati, Amin. (2020). Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree. Reliab. Eng. Syst. Saf., 200, 106931. http://doi.org/10.1016/j.ress.2020.106931
26. Zhu, Min., Xia, Jing., Jin, Xiaoqing., Yan, M., Cai, Guolong., Yan, Jing., & Ning, Gangmin. (2018). Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. IEEE Access, 6, 4641-4652. http://doi.org/10.1109/ACCESS.2018.2789428