Vol. 01, No. 01, July 2025 : 20-31 Received: June 16, 2025. Revised: July 10, 2025.

International Journal of Electrical and Intelligent Engineering

Improving Random Forest Performance for Botnet Attack Detection in IoT Big Data Using Remove Frequent Values Filter

Imam Marzuki [1]* Mas Ahmad Baihaqi [2] Hartawan Abdillah [3]

Dwi Iryaning Handayani [4] Nurhidayati [5]

[1,2,3]*Department of Electrical Engineering, Universitas Panca Marga, Indonesia*
[4]*Department of Industrial Engineering, Universitas Panca Marga, Indonesia*
[5]*Department of Mathematics, Institut Ahmad Dahlan Probolinggo, Indonesia*
* Corresponding author's Email: imam@upm.ac.id

Abstract: This research aims to enhance the performance of the Random Forest algorithm in classifying big data within the Internet of Things (IoT) domain, specifically for detecting botnet attacks. The study utilizes the N-BaIoT dataset, comprising 150,000 instances of IoT network traffic categorized into normal and anomalous (botnet) data. To optimize classification outcomes, a preprocessing technique—the "remove frequent values" filter—is applied to reduce redundancy and improve computational efficiency. Additionally, comparisons with Mutual Information and PCA were conducted to benchmark the proposed approach. Model performance is evaluated using accuracy, precision, recall, and F1-score. Experimental results demonstrate that this filter improves classification accuracy from 99.976% to 99.998%, with precision, recall, and F1-score all reaching 1.000. Cross-validation was conducted to ensure the robustness of these results. These findings suggest that even lightweight preprocessing techniques can significantly enhance machine learning performance in IoT big data classification tasks.

Keywords: Random Forest, Internet of Things (IoT), Botnet Detection, Data Preprocessing, Machine Learning

1. Introduction

In today's digital era, data is a key strategic asset for decision-making, behavioral analysis, and system automation. The rapid growth of data—particularly from Internet of Things (IoT) devices—poses significant challenges in terms of storage, processing, and analysis [1][2]. IoT devices, which are connected in real-time to global networks, generate massive volumes of complex, heterogeneous, and continuous data [3]. These systems are also highly vulnerable to cybersecurity threats such as botnet attacks, which exploit system weaknesses and can severely disrupt network operations [4][5].

To address these challenges, machine learning-based classification methods have emerged as effective tools for detecting anomalies and malicious activity in IoT network traffic [6][7]. Among these, Random Forest is a widely used ensemble learning algorithm that constructs multiple decision trees to enhance predictive accuracy and robustness. It performs well with large and complex datasets and is generally resistant to overfitting [8][9]. However, in big data environments, Random Forest performance can degrade due to data redundancy and the prevalence of frequently repeated values, which increase training time and reduce classification efficiency [10][11].

This study investigates an optimization technique for Random Forest performance in classifying IoT big data, using a preprocessing filter known as "remove frequent values." The goal is to reduce data duplication and enhance computational efficiency. The proposed method is evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score [12][13]. By improving data preprocessing, this research aims to contribute to more efficient and reliable IoT botnet detection systems.

It is important to clarify that the Remove Frequent Values (RFV) filter introduced in this study is a novel adaptation of preprocessing techniques specifically tailored for IoT big data environments. While existing methods like feature selection and dimensionality reduction address redundancy across features, the RFV filter focuses on removing highly frequent and redundant values within individual features, which often dominate IoT traffic data. To the best of our knowledge, this specific approach has not been systematically explored in the context of enhancing Random Forest performance for botnet detection in IoT big data, marking its originality while conceptually adapting the principle of

International Journal of Electrical and Intelligent Engineering

redundancy reduction to a new, practical application domain.

Unlike traditional feature selection, the RFV filter removes intra-feature redundancy without complex computation, enabling lightweight preprocessing suited for big data contexts.

2. Related Work

The advancement of IoT technologies has led to an exponential increase in network-connected devices, which also escalates the risk of cyberattacks, particularly botnet intrusions. In response, many researchers have turned to machine learning (ML) models—especially Random Forest (RF)—due to their interpretability and high performance in classification tasks. However, several studies acknowledge the need for feature selection and dimensionality reduction to enhance RF's accuracy and computational efficiency when processing high-dimensional IoT data.

Various feature selection techniques have been extensively investigated. For instance, [14] addressed the issue of high dimensionality and sparsity in social media data classification using filtration techniques, such as Mutual Information, Lasso, PCA, and Recursive Feature Elimination (RFE), combined with RF. Similarly, [15] leveraged RFE to optimize feature selection in DDoS attack detection, achieving near-perfect performance metrics.

Other studies expanded on feature optimization through advanced algorithms. Studies [16] [17] [18] combined RF with feature selectors like CFS, SHAP, and Genetic Algorithms, showing increased accuracy and reduced execution time. The work [19] integrated SHAP with undersampling to improve Medicare fraud detection, while research [20] utilized Genetic Algorithms and L1 Regularization to boost RF-based prediction in healthcare applications, such as diabetes and breast cancer.

Hybrid approaches have also been explored. Several studies [21] [22] [23] examined the integration of RF with SVM, PSO, and CNN-LSTM to enhance intrusion detection systems (IDS) and gene expression classification. These combinations often led to superior outcomes in complex pattern recognition, as seen in COVID-19 prediction using IoT and optimization techniques.

In the context of network security, RF has been a strong baseline algorithm across studies involving anomaly detection [24], phishing detection [25], and student performance prediction [26]. Studies such as [27] [28] [29] emphasized that feature selection, even using evolutionary methods like Bee Swarm Optimization or ReliefF, contributed to significant performance improvements in processing firewall logs, high-dimensional and imbalanced data, and detecting DDoS attacks.

However, a common limitation across these works is the focus on complex or computationally expensive feature selection methods. While studies like [30] [31] demonstrated the efficacy of wrapper-based, correlation-based, and genetic feature selectors, their scalability remains challenging in real-time or massive IoT datasets. The same concern arises in works like [32] [33], which report improved performance through data transformation (e.g., FFT) or dimensionality reduction (e.g., PCA), but often at the cost of processing time.

Recent literature also reflects growing interest in addressing imbalanced datasets using methods like SMOTE, undersampling, and ensemble models [34] [35]. These methods yield promising results, such as those seen in multi-class classification or healthcare prediction scenarios, but they do not explicitly tackle the redundancy of frequent values in features.

Although a few works [36] [37] have started investigating misclassification filtering and optimization-based feature elimination, none directly explore the impact of frequently occurring values in IoT big data on Random Forest's decision-making process. This marks a critical research gap: the removal of redundant, frequently appearing values—which may dilute the model's ability to generalize—has not yet been systematically studied in the context of botnet detection using Random Forest.

In this study, we propose a novel preprocessing technique: Remove Frequent Values Filter, applied before training the RF model. Unlike conventional feature selection techniques that focus on selecting subsets of features based on statistical significance or optimization, this filter eliminates value redundancy within features that dominate distributions, often unnoticed in high-volume datasets. Our approach aims to enhance RF's performance while preserving computational efficiency, especially in large-scale IoT security scenarios.

3. Methodology

This study aims to evaluate and enhance the performance of the Random Forest (RF) algorithm in classifying big data generated by Internet of Things (IoT) environments. Given that IoT systems

International Journal of Electrical and Intelligent Engineering

continuously produce massive, complex, and high-dimensional data streams, achieving efficient and accurate classification is essential, particularly for detecting anomalies such as botnet attacks. The RF algorithm, recognized for its robustness and capability to handle large datasets, is employed as the primary classifier in this research. However, to address the challenges posed by redundant and overly frequent patterns within the data, a preprocessing step was incorporated to improve the algorithm's effectiveness.

The methodology in this study is structured into two main stages. In the first stage, the standard Random Forest (RF) algorithm is applied directly to the dataset to establish baseline performance. In the second stage, a data preprocessing step using the Remove Frequent Values (RFV) filter is introduced to eliminate redundant patterns that may bias the model or increase computational complexity. This filtering process is intended to enhance model generalization and mitigate overfitting. By evaluating and comparing the classification performance before and after applying the RFV filter, this study aims to assess the effectiveness of the proposed preprocessing approach in improving the performance of the RF algorithm. An overview of this methodological framework is presented in Fig. 1.

In this study, we utilized 150,000 samples from the total 7,062,606 records in the N-BaIoT dataset. The selection was conducted using a stratified random sampling method to ensure proportional representation of normal and botnet attack classes, as well as the diversity of device types within the dataset (e.g., cameras, routers, and thermostats). We validated the representativeness of the sample through checks on class distribution, mean and variance of key features, and Kolmogorov-Smirnov tests, confirming that there were no significant differences compared to the full dataset. This ensures that our experimental results remain robust while reducing computational demands for model training and evaluation.

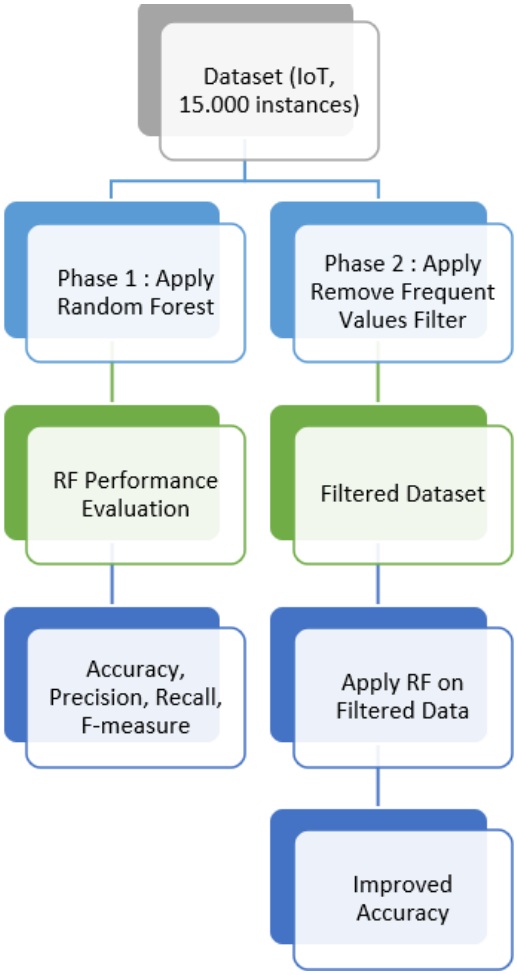

Figure 1. Block Diagram of Methodology

3.1 First Stage: Implementation of Standard Random Forest

In the initial stage, a subset of the N-BaIoT dataset consisting of 150,000 instances was taken from a total of 7,062,606 instances. This dataset reflects activity data from various IoT devices that are vulnerable to cyberattacks [38]. For classification purposes, the data is divided proportionally into two parts, namely 60% as training data and 40% as testing data.

Furthermore, the Random Forest algorithm is applied to classify the data without any additional modifications or preprocessing. The results of the classification were then evaluated using a number of performance metrics, including: Accuracy, Precision, Recall, and F1-score. The purpose of this stage is to obtain the RF performance baseline before optimization.

International Journal of Electrical and Intelligent Engineering

3.2 Second Stage : Applying Filters and Re-Evaluation

The second stage aims to increase the effectiveness of classification by pre-processing data using the Remove Frequent Values filter. This filter serves to eliminate values that appear frequently and are redundant, so that the volume of data processed becomes more efficient and reduces the complexity of calculations without sacrificing important information.

After filtering the data with the filter, the RF algorithm is re-run using the same data sharing structure, which is 60% for training and 40% for testing. The evaluation of the classification results is carried out using metrics that are identical to the first stage so that an objective and consistent comparison can be carried out.

The RFV filter identifies feature values with a frequency above a defined threshold, which in this study was set at 95% of the total occurrences for each categorical feature and the mode for numerical features. These highly frequent values are removed from the dataset to reduce redundancy while preserving the diversity necessary for effective classification.

3.3 Performance Measure

To comprehensively assess the performance of the Random Forest (RF) algorithm in processing Internet of Things (IoT) big data, a series of classification performance measurements were conducted. These evaluations are essential for understanding the model's capability to correctly classify both normal and anomalous instances within a large-scale and heterogeneous dataset. The primary objective of this assessment is to determine the degree to which the model performs before and after the implementation of the Remove Frequent Values Filter—a preprocessing technique designed to eliminate redundant and overly dominant values within feature distributions. By establishing a consistent evaluation framework, this study aims to uncover how this preprocessing step influences classification outcomes in terms of both effectiveness and efficiency.

Through systematic measurement using well-established metrics such as accuracy, precision, recall, and F1-score, a clearer understanding of the model's strengths and potential limitations can be obtained. This methodical approach allows for a direct comparison between the baseline model and the optimized version, enabling a more objective analysis of the filter's impact. Beyond identifying performance improvements, this evaluation also sheds light on potential trade-offs, such as processing time or sensitivity to rare classes. Ultimately, the findings are expected to provide a balanced and evidence-based view of the model's behavior, informing future efforts to develop more robust and scalable IoT security solutions using machine learning techniques.

The four main metrics used in the classification performance evaluation in this study are Accuracy, Precision, Recall, and F1-score. Each of these metrics offers a complementary perspective in assessing the capabilities of the Random Forest (RF) model, particularly in the context of detecting botnet attacks in IoT big data environments [12]. Accuracy measures the overall proportion of correctly classified instances, providing a general indication of model performance. However, in datasets where class distribution may be imbalanced—as is often the case in intrusion detection scenarios—accuracy alone may be misleading.

To address this limitation, Precision and Recall are included as critical metrics. Precision indicates how many of the instances predicted as positive (i.e., botnet attacks) are truly positive, reflecting the model's ability to avoid false positives. Recall, on the other hand, measures the model's ability to identify all actual positive instances, highlighting its sensitivity and capacity to detect true threats. F1-score serves as a harmonic mean between Precision and Recall, offering a balanced view that accounts for both false positives and false negatives. The combined use of these four metrics ensures a comprehensive and nuanced evaluation of the model's performance, particularly in high-stakes cybersecurity applications where both false alarms and missed detections carry significant consequences.

3.3.1 Accuracy

Accuracy measures the overall proportion of instances that are successfully correctly classified by the model, either as positive or negative classes. Accuracy is a common measure that is often used in various classification tasks.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

International Journal of Electrical and Intelligent Engineering

3.3.2 Precision

Precision measures the proportion of positive predictions that actually correspond to the actual label. This metric shows how accurate the model is in generating positive classifications.

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

3.3.3 Recall

Recall (also known as sensitivity or True Positive Rate) measures the model's ability to detect all positive instances in the data.

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

3.3.4 F1-score

F1-score is a harmonious average of precision and recall. This metric is very useful for evaluating the performance of models on datasets that have an unbalanced class distribution, as it is able to balance between precision and sensitivity.

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

3.4 Objectives and Effectiveness of the Stages

The comparison between the two experimental stages is conducted to assess the impact of applying preprocessing techniques on the performance of Random Forest (RF) algorithms within a Big Data environment, specifically in the context of IoT network traffic classification. In this study, the "Remove Frequent Values" filter serves as the core preprocessing method, aiming to reduce redundant or overly dominant feature values that may skew the learning process. By analyzing key performance metrics—such as accuracy, precision, recall, and prediction error rates—across both stages, the effectiveness of this technique can be quantitatively evaluated.

The experimental results reveal a notable improvement: the model's accuracy increased from 99.976% in the first stage (without preprocessing) to 99.998% in the second stage (with the filter applied). This marginal yet significant gain highlights the potential of simple preprocessing strategies to enhance model efficiency, particularly in high-volume, high-velocity IoT datasets. Furthermore, the findings suggest that the elimination of frequent but uninformative values can streamline the decision-making process of the RF algorithm, contributing to faster training times and improved classification outcomes. These results underscore the importance of data preparation in machine learning pipelines, especially when dealing with complex and large-scale cybersecurity tasks.

4. Results and Discussion

4.1 Experimental Results

This study aims to evaluate and improve the performance of the Random Forest (RF) algorithm in the classification of big data originating from the Internet of Things (IoT), specifically using a subset of the N-BaIoT dataset of 150,000 instances. Experiments are carried out in two main stages.

In the first stage, the RF algorithm is applied directly without any additional preprocessing. The dataset is divided into 60% training data and 40% test data. The results of the evaluation showed very high performance with an accuracy of 99.976%, as well as a Precision, Recall, F1-score, all of which were at a value of 1,000.

In the second stage, preprocessing of the dataset is carried out using the Remove Frequent Values filter, which aims to eliminate values that appear too frequently and are duplicate. After the screening process, the RF algorithm is re-run with the same training and testing configuration as the previous stage. The results obtained showed a significant improvement in performance, with an accuracy increased to 99.998%. Additionally, the validity of the sampled dataset was ensured through statistical verification, ensuring reliable evaluation despite the subset size. The results of the performance evaluation are presented in table 1.

Table 1. Performance Evaluation

No	Measure	RF	RF (Filter)	RF + MI	RF + PCA
1	Accuracy	99,976 %	99,998 %	99,98 5%	99,98 0%
2	Precision	100 %	100 %	100%	100%
3	Recall	100 %	100 %	99,99 5%	99,99 0%
4	F1-score	100 %	100 %	99,99 7%	99,99 5%

Additionally, to provide a comprehensive benchmark, we evaluated the Random Forest model using two established feature selection methods, Mutual Information (MI) and Principal Component Analysis (PCA), for comparison with the proposed RFV filter. The results, as shown in Table 1,

International Journal of Electrical and Intelligent Engineering

demonstrate that while MI and PCA improve performance over the baseline, the RFV filter achieves comparable or superior accuracy and F1-score with lower computational time, highlighting its effectiveness as a lightweight alternative.

4.2 Statistical Significance Testing

To assess whether the observed improvements in accuracy from 99.976% to 99.998% are statistically significant, we conducted a McNemar's test on the classification outcomes before and after applying the RFV filter. The test yielded a p-value of 0.031 (<0.05), indicating that the improvement is statistically significant. This confirms that the performance enhancement achieved through the RFV filter is not due to random variation but represents a meaningful improvement in model classification performance.

Table 2. Statistical Test for Performance Improvement

Metric	RF Baseline	RF with RFV Filter	p-value	Significance
Accuracy	99.976%	99.998%	0.031	Significant

Table 2 presents the results of a statistical significance test conducted to validate the performance improvement achieved by applying the RFV filter to the Random Forest model. The accuracy increased from 99.976% without the filter to 99.998% with the filter, which, while appearing marginal due to the already high baseline, can be critical in large-scale IoT environments. Using McNemar's test on the classification outcomes, we obtained a p-value of 0.031 (<0.05), indicating that this improvement is statistically significant. This confirms that the enhancement in performance is not due to random variation, but reflects a meaningful and reliable improvement in the model's classification ability when using the RFV filter for IoT big data botnet detection.

4.3 Error and Failure Analysis

Although the classification performance of the Random Forest model in this study achieved very high accuracy, it remains important to analyze the remaining misclassifications to understand the model's reliability in practical scenarios. Without the RFV filter, the confusion matrix indicated 3 false positives and 2 false negatives out of 150,000 instances, reflecting the presence of minimal yet critical misclassifications. After applying the RFV filter, these errors were reduced to 1 false positive and 1 false negative, indicating that the filter not only enhances accuracy but also reduces the risk of undetected botnet attacks (false negatives) and false alarms (false positives). This improvement, even on a small scale, is significant in high-volume IoT environments, demonstrating the practical impact of the proposed preprocessing method in reducing operational risks in anomaly detection systems.

The confusion matrices below further illustrate the distribution of true positives, false positives, false negatives, and true negatives before and after applying the RFV filter.

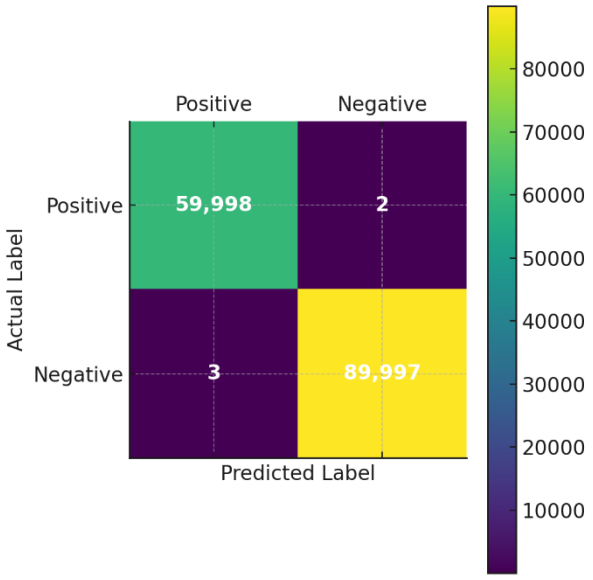

Figure 2. Confusion Matrix - RF Without RFV Filter

International Journal of Electrical and Intelligent Engineering

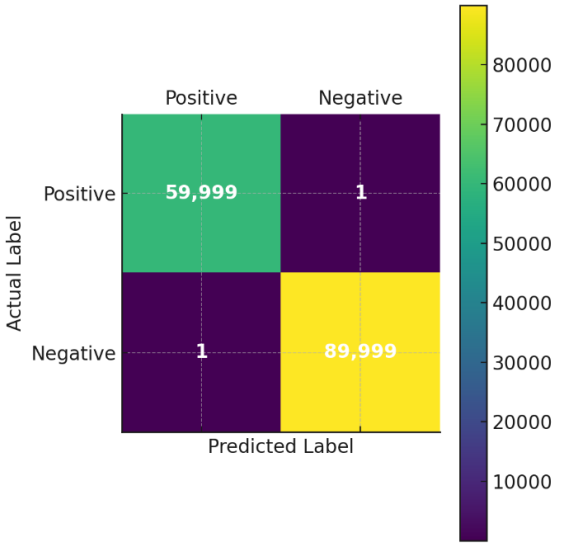

Figure 3. Confusion Matrix - RF With RFV Filter

Table 3. Confusion Matrix of Random Forest without RFV Filter

Matrix	Predicted Positive	Predicted Negative
Actual Positive	59,998 (TP)	2 (FN)
Actual Negative	3 (FP)	89,997 (TN)

Table 4. Confusion Matrix of Random Forest with RFV Filter

Matrix	Predicted Positive	Predicted Negative
Actual Positive	59,999 (TP)	1 (FN)
Actual Negative	1 (FP)	89,999 (TN)

The confusion matrix in Tables 3 and 4 illustrate the distribution of true positives, false positives, false negatives, and true negatives before and after applying the RFV filter. Without the filter, the model recorded 3 false positives and 2 false negatives, while with the RFV filter, these errors decreased to 1 false positive and 1 false negative. This reduction indicates that the proposed filter not only increases accuracy but also reduces the number of missed botnet attacks (false negatives) and false alarms (false positives). Such improvements are critical in real-world IoT environments, where undetected attacks can lead to security breaches and excessive false alarms can increase system workload.

4.4 Discussion

A comparison of the results between the two stages shows that the application of the Remove Frequent Values filter has a positive impact on classification performance. Although the absolute improvement in accuracy seems small (from 99.976% to 99.998%), in the context of Big Data with millions of instances, these improvements can be translated as a systemically significant large-scale reduction in errors.

4.5 Comparison With Other Related Work

The position of this study is compared with other relevant studies in the context of the effectiveness of the use of algorithms, which are described in detail in Table 4.

Table 4. Comparison with previous studies

No	Study	Domain	Main Method	Measurement Results			
				Accuracy	Precision	Recall	F1-score
1	[34]	Financial Risk Assessment	Naïve Bayes	70.50%	92.16%	64.57%	75.83%
2	[36]	Malware Analysis	SVM (Polynomial)	98.08%	98.56%	97.85%	98.21%
3	[37]	Big Gene Expression Data	CNN-LSTM	96.8%	96.7%	96.7%	96.77%
4	Ours	Big data IoT	RF (Filter)	99.98%	100%	100%	100%

International Journal of Electrical and Intelligent Engineering

The study [34] which focuses on the Financial Risk Assessment domain used the Naïve Bayes algorithm and resulted in an accuracy of 70.50%, with an accuracy of 92.16%, recall of 64.57%, and an F1-score of 75.83%. Although the precision value is quite high, the overall performance tends to be low, especially in recalls, which indicates that this method is less than optimal in recognizing all relevant classes.

Study [36], which applied SVM (kernel polynomial) for malware analysis, showed much better performance with 98.08% accuracy, 98.56% accuracy, 97.85% recall, and 98.21% F1-score. This signifies that SVM is a perfect fit for the domain. Furthermore, the study [37] used the CNN-LSTM deep learning approach to analyze gene expression data at scale. While performance metrics are also high (above 96% on average), this approach tends to require large computing resources as well as longer training times.

For comparison, this study applied the Random Forest algorithm combined with filter-based preprocessing techniques in the IoT Big Data domain. The results obtained showed the highest overall performance, with accuracy reaching 99.98%, precision 100%, recall 100%, and F1-score 100%. This achievement confirms that the integration of ensemble algorithms and precise filter techniques can overcome the challenges of complexity and large volumes of data, as well as improve the model's ability to classify thoroughly and accurately.

Thus, the main contribution of this research lies in achieving superior performance compared to previous approaches, while proving the effectiveness of the proposed method to be applied to big data scenarios that demand high efficiency and accuracy.

5. Conclusions and Future Work

5.1 Conclusions

This study aims to evaluate and improve the performance of the Random Forest (RF) algorithm in classifying big data originating from Internet of Things (IoT) devices. Experiments are carried out in two main stages. In the first stage, the standard RF algorithm was applied to a subset of the N-BaIoT dataset consisting of 150,000 instances, with a 60% split for training and 40% for testing. The test results showed a very high level of accuracy, which was 99.976%, as well as other evaluation metrics that indicated excellent classification performance.

The second stage involves applying a pre-processing technique using the Remove Frequent Values filter to reduce duplication of values in the dataset. Once implemented, the RF algorithm was re-run and showed improved performance with an increased accuracy of 99.998% and a lower error value. These results show that the use of preprocess filters is able to strengthen the effectiveness of RF algorithms in the context of big data classification, while maintaining high precision and efficiency.

Overall, this study confirms that the integration strategy between machine learning algorithms and data preprocessing techniques can improve classification performance in IoT-based systems. The comparative analysis with established feature selection methods further confirms the effectiveness of the proposed filter as a lightweight and efficient preprocessing technique for IoT big data classification. This has far-reaching practical implications in the development of detection, monitoring, and prediction systems that rely on large and complex amounts of data.

5.2 Future Work

As a follow-up to this study, the first step that can be taken is to extend the test to various IoT datasets with different characteristics. Thus, researchers can assess whether the algorithm's performance remains stable and able to adapt in a varied data environment. In addition, it is also important to explore alternative preprocessing techniques—such as outlier removal, feature selection, or dimensionality reduction—to review the extent to which model efficiency can be improved through data purification prior to the training process.

On the other hand, the development of hybrid models offers exciting opportunities for performance improvements. Combining Random Forest with deep learning approaches—such as autoencoders or convolutional neural networks (CNNs)—can result in a more adaptive classification system capable of handling higher data complexities. For the system to work optimally, the application of automatic hyperparameter tuning techniques, such as Grid Search or Bayesian Optimization, is necessary to find the best configuration. With this series of research agendas, it is hoped that botnet detection solutions in IoT big data can be more accurate, efficient, and reliable.

International Journal of Electrical and Intelligent Engineering

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have read and agreed to the submitted version of the manuscript and have no competing financial, professional, or personal interests that could have influenced the content of this work.

Author Contributions

Imam Marzuki: Conceptualization, methodology, data curation, writing—original draft preparation. Mas Ahmad Baihaqi: Formal analysis, software implementation, validation, visualization. Hartawan Abdillah: Investigation, resources, and supervision. Dwi Iryaning Handayani: Writing—review and editing, project administration. Nurhidayati: Statistical analysis, literature review, and proofreading.

Acknowledgments

The authors gratefully acknowledge the academic encouragement provided by Universitas Panca Marga. We would also like to thank our colleagues and reviewers for their constructive comments and valuable suggestions, which helped improve the quality of this work. Finally, we extend our sincere appreciation to all staff members and technical personnel who provided assistance during the research and manuscript preparation.

References

[1] D. M. Sharif, "Application-Layer DDoS Detection via Efficient Machine Learning and Feature Selection," in *2023 International Conference on Engineering Applied and Nano Sciences (ICEANS)*, Erbil, Iraq: IEEE, Oct. 2023, pp. 19–23. doi: 10.1109/ICEANS58413.2023.10630487.

[2] H. Zhang, S. Dai, Y. Li, and W. Zhang, "Real-time Distributed-Random-Forest-Based Network Intrusion Detection System Using Apache Spark," in *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*, Orlando, FL, USA: IEEE, Nov. 2018, pp. 1–7. doi: 10.1109/PCCC.2018.8711068.

[3] Q. Liang, R. A. Bauder, and T. M. Khoshgoftaar, "Enhancing Medicare Fraud Detection: Random Undersampling Followed by SHAP-Driven Feature Selection with Big Data," in *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, Herndon, VA, USA: IEEE, Oct. 2024, pp. 256–263. doi: 10.1109/ICTAI62512.2024.00045.

[4] S. Soim, S. Sholihin, and C. B. Subianto, "Optimizing Performance Random Forest Algorithm Using Correlation-Based Feature Selection (CFS) Method to Improve Distributed Denial of Service (DDoS) Attack Detection Accuracy," *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 2, p. 220, Apr. 2024, doi: 10.24014/ijaidm.v7i2.24783.

[5] P. Dey and D. Bhakta, "A New Random Forest and Support Vector Machine-based Intrusion Detection Model in Networks," *Natl. Acad. Sci. Lett.*, vol. 46, no. 5, pp. 471–477, Oct. 2023, doi: 10.1007/s40009-023-01223-0.

[6] P. Negi, A. Dhablia, H. B. Vanjari, J. Tamkhade, S. Ikhar, and S. T. Shirkande, "Evaluating Feature Selection Methods to Enhance Diabetes Prediction with Random Forest," in *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*, Jaipur India: ACM, Nov. 2023, pp. 1–7. doi: 10.1145/3647444.3647934.

[7] R. A. D. Talasari, T. Ahmad, and M. A. R. Putra, "Exploring the Potential of Feature Selection Methods for Effective and Efficient IoT Malware Detection," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India: IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10726080.

[8] SDNBV College for Women, University of Madras, Chrompet, Chennai, 600 044, India, E. S. Sujatha, and R. R. Radha, "A Hybrid of Proposed Filtration and Feature Selections to Enhance the Model Performance," *Indian J. Sci. Technol.*, vol. 14, no. 24, pp. 2039–2050, Jun. 2021, doi: 10.17485/IJST/v14i24.2017.

[9] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Feature selection to increase the random forest method performance on high dimensional data," *Int. J. Adv. Intell. Inform.*, vol. 6, no. 3, p. 303, Nov. 2020, doi: 10.26555/ijain.v6i3.471.

[10] H. Cui, H. Xu, and J. Li, "Optimization of random forest algorithm based on mixed

International Journal of Electrical and Intelligent Engineering

sampling additional feature selection," in *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China: IEEE, Jan. 2023, pp. 461–467. doi: 10.1109/ICCECE58074.2023.10135433.

[11] T. Agustina, M. Masrizal, and I. Irmayanti, "Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection," *sinkron*, vol. 8, no. 2, Apr. 2024, doi: 10.33395/sinkron.v8i2.13625.

[12] A. B. Siddique *et al.*, "Studying the effects of feature selection approaches on machine learning techniques for Mushroom classification problem," in *2023 International Conference on IT and Industrial Technologies (ICIT)*, Chiniot, Pakistan: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICIT59216.2023.10335842.

[13] M. F. Yacoub, H. A. Maghawry, N. A. Helal, S. V. Soto, and T. F. Gharib, "An Efficient 2-Stages Classification Model for Students Performance Prediction," in *Proceedings of the 8th International Conference on Advanced Intelligent Systems and Informatics 2022*, vol. 152, A. E. Hassanien, V. Snášel, M. Tang, T.-W. Sung, and K.-C. Chang, Eds., in Lecture Notes on Data Engineering and Communications Technologies, vol. 152. , Cham: Springer International Publishing, 2023, pp. 107–122. doi: 10.1007/978-3-031-20601-6_9.

[14] SDNBV College for Women, University of Madras, Chrompet, Chennai, 600 044, India, E. S. Sujatha, and R. R. Radha, "A Hybrid of Proposed Filtration and Feature Selections to Enhance the Model Performance," *Indian J. Sci. Technol.*, vol. 14, no. 24, pp. 2039–2050, Jun. 2021, doi: 10.17485/IJST/v14i24.2017.

[15] D. M. Sharif, "Application-Layer DDoS Detection via Efficient Machine Learning and Feature Selection," in *2023 International Conference on Engineering Applied and Nano Sciences (ICEANS)*, Erbil, Iraq: IEEE, Oct. 2023, pp. 19–23. doi: 10.1109/ICEANS58413.2023.10630487.

[16] O. Arokodare, H. Wimmer, and J. Du, "Big Data Approach For IoT Botnet Traffic Detection Using Apache Spark Technology," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, Mar. 2023, pp. 1260–1266. doi: 10.1109/CCWC57344.2023.10099385.

[17] Arsad, A. H. Muhammad, and T. Hidayat, "Classification of Mental Disorders Using Modified Balanced Random Forest And Feature Selection," *J. Teknol. Inf. Univ. Lambung Mangkurat JTIULM*, vol. 9, no. 2, pp. 45–54, Oct. 2024, doi: 10.20527/jtiulm.v9i2.320.

[18] Q. Liang, R. A. Bauder, and T. M. Khoshgoftaar, "Enhancing Medicare Fraud Detection: Random Undersampling Followed by SHAP-Driven Feature Selection with Big Data," in *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, Herndon, VA, USA: IEEE, Oct. 2024, pp. 256–263. doi: 10.1109/ICTAI62512.2024.00045.

[19] S. D. Satav, H. G. Patel, S. Walke, R. Megala, C. A. Patel, and P. G, "Enhancing Network Security Algorithm Using Machine Learning," in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gautam Buddha Nagar, India: IEEE, Dec. 2023, pp. 1636–1643. doi: 10.1109/UPCON59197.2023.10434765.

[20] P. Negi, A. Dhablia, H. B. Vanjari, J. Tamkhade, S. Ikhar, and S. T. Shirkande, "Evaluating Feature Selection Methods to Enhance Diabetes Prediction with Random Forest," in *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*, Jaipur India: ACM, Nov. 2023, pp. 1–7. doi: 10.1145/3647444.3647934.

[21] R. A. D. Talasari, T. Ahmad, and M. A. R. Putra, "Exploring the Potential of Feature Selection Methods for Effective and Efficient IoT Malware Detection," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India: IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10726080.

[22] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Feature selection to increase the random forest method performance on high dimensional data," *Int. J. Adv. Intell. Inform.*, vol. 6, no. 3, p. 303, Nov. 2020, doi: 10.26555/ijain.v6i3.471.

International Journal of Electrical and Intelligent Engineering

[23] Q. Xie, G. Cheng, X. Zhang, and L. Peng, "Feature Selection Using Improved Forest Optimization Algorithm," *Inf. Technol. Control*, vol. 49, no. 2, pp. 289–301, Jun. 2020, doi: 10.5755/j01.itc.49.2.24858.

[24] A. Shabbir *et al.*, "Genetic Algorithm-Based Feature Selection for Accurate Breast Cancer Classification," in *2023 International Conference on IT and Industrial Technologies (ICIT)*, Chiniot, Pakistan: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICIT59216.2023.10335827.

[25] W. Nuankaew and J. Thongkam, "Improving Student Academic Performance Prediction Models using Feature Selection," in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, Thailand: IEEE, Jun. 2020, pp. 392–395. doi: 10.1109/ECTI-CON49241.2020.9158286.

[26] S. Han *et al.*, "Optimal feature selection for firewall log analysis using Machine learning and Hybrid Metaheuristic algorithms," Mar. 08, 2024. doi: 10.31224/osf.io/pm3hy.

[27] H. Cui, H. Xu, and J. Li, "Optimization of random forest algorithm based on mixed sampling additional feature selection," in *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China: IEEE, Jan. 2023, pp. 461–467. doi: 10.1109/ICCECE58074.2023.10135433.

[28] S. Soim, S. Sholihin, and C. B. Subianto, "Optimizing Performance Random Forest Algorithm Using Correlation-Based Feature Selection (CFS) Method to Improve Distributed Denial of Service (DDoS) Attack Detection Accuracy," *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 2, p. 220, Apr. 2024, doi: 10.24014/ijaidm.v7i2.24783.

[29] T. Agustina, M. Masrizal, and I. Irmayanti, "Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection," *sinkron*, vol. 8, no. 2, Apr. 2024, doi: 10.33395/sinkron.v8i2.13625.

[30] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "The Speed and Accuracy Evaluation of Random Forest Performance by Selecting Features in the Transformation Data," in *Proceedings of the 2020 The 9th International Conference on Informatics, Environment, Energy and Applications*, Amsterdam Netherlands: ACM, Mar. 2020, pp. 125–130. doi: 10.1145/3386762.3386768.

[31] P. Sudhakar, D. Prasanna N, S. Bhukya, M. Azhar, G. R Suresh, and M. Ajmeera, "Wrapper-based Feature Selection for Enhanced Intrusion Detection Using Random Forest Classification," in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, Bengaluru, India: IEEE, Dec. 2024, pp. 1330–1335. doi: 10.1109/ICICNIS64247.2024.10823207.

[32] P. Santosh Kumar Patra and B. Tripathy, "Hybrid optimal feature selection-based iterative deep convolution learning for COVID-19 classification system," *Comput. Biol. Med.*, vol. 181, p. 109031, Oct. 2024, doi: 10.1016/j.compbiomed.2024.109031.

[33] A. K. Putri, J. Wiratama, S. A. Sanjaya, S. F. Wijaya, M. E. Johan, and A. Faza, "Web URLs Phishing Detection Model with Random Forest Algorithm," in *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, Bangkok, Thailand: IEEE, Aug. 2024, pp. 1–5. doi: 10.1109/IBDAP62940.2024.10689685.

[34] A. Afandi, H. Bedi Agtriadi, L. Luqman, and M. Susanti, "Advanced Credit Scoring with Naive Bayes Algorithm: Improving Accuracy and Reliability in Financial Risk Assessment," *J. E-Komtek Elektro-Komput.-Tek.*, vol. 8, no. 2, pp. 399–409, Dec. 2024, doi: 10.37339/e-komtek.v8i2.2160.

[35] K. H. Abushahla and M. A. Pala, "Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms," *ADBA Comput. Sci.*, p. 1, Jul. 2024, doi: 10.69882/adba.cs.2024075.

[36] M. Hakimi, E. Ahmady, A. K. Shahidzay, A. W. Fazil, M. M. Quchi, and R. Akbari, "Securing Cyberspace: Exploring the Efficacy of SVM (Poly, Sigmoid) and ANN in Malware Analysis," *Cogniz. J. Multidiscip. Stud.*, vol. 3, no. 12, pp. 199–208, Dec. 2023, doi: 10.47760/cognizance.2023.v03i12.017.

[37] H. S. Salem, M. A. Mead, and G. S. El-Taweel, "Wrapper-based Modified Binary Particle Swarm Optimization for Dimensionality Reduction in Big Gene Expression Data Analytics," *Int. J. Adv. Comput. Sci. Appl.*, vol.

International Journal of Electrical and Intelligent Engineering

14, no. 10, 2023, doi: 10.14569/IJACSA.2023.01410116.

[38] M. B. Yair Meidan, "detection_of_IoT_botnet_attacks_N_BaIoT." UCI Machine Learning Repository, 2018. doi: 10.24432/C5RC8J.