Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING MALWARE DETECTION USING HYBRID FEATURE SELECTION TECHNIQUES IN PREDICTIVE MODELS

GAYATHRI DEVI N[1], V. KRISHNA[2], NEELIMA GURRAPU[3], RAJESH BANALA[4], SHAIK JILANI BASHA[5], KOMATI SATHISH[6]

[1]Associate Professor, Department of IT, C. Abdul Hakeem College of Engineering and Technology, Hakeem Nagar, Melvisharam, India
[2]Associate Professor, Department of CSE, TKR College of Engineering and Technology, Hyderabad, Telangana, India
[3]Assistant Professor, Department of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India
[4,6]Assistant Professor, Department of CSE, TKR College of Engineering and Technology, Hyderabad, Telangana, India
[5]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India

Email: [1]gayathridevi.nphd2017@gmail.com, [2]vempati.k@gmail.com, [3]gneelima83@gmail.com, [4]rajesh.banala@gmail.com, [5]jilani.1221@gmail.com, [6]komatisathish459@gmail.com

ABSTRACT

Malware detection remains a critical challenge in cybersecurity due to the growing complexity and diversity of malicious threats. Existing literature largely focuses on either filter-based or wrapper-based feature selection methods, often limited to specific malware categories or datasets. This study addresses this gap by introducing a hybrid feature selection approach that integrates statistical filter metrics with model-specific wrapper refinement, aiming to optimize feature subsets for enhanced malware detection. Publicly available datasets, including the Microsoft Malware Classification Challenge dataset and the Kaggle Malware Dataset, were used to evaluate multiple models such as Gradient Boosting and Neural Networks. Experimental results show that the hybrid approach consistently outperforms individual methods in terms of accuracy, recall, and computational efficiency, achieving up to 95% accuracy and 96% ROC-AUC. The findings contribute new knowledge by presenting a generalizable, scalable framework that improves malware detection performance across heterogeneous features and datasets.

Keywords: *Malware Detection, Hybrid Feature Selection, Machine Learning, Filter And Wrapper Techniques, Cyber Security*

1. INTRODUCTION

Malware poses a significant threat to information security, causing extensive financial and reputational damage to individuals and organizations worldwide. To effectively detect malware, leveraging advanced machine learning algorithms is essential, as traditional signature-based methods struggle to keep pace with the growing complexity of cyber threats. In these machine learning models, feature selection plays a critical role by identifying the most relevant variables from large datasets, thereby enhancing model accuracy and reducing computational demands. Previous research in credit card fraud detection has demonstrated the efficacy of hybrid feature selection techniques, which integrate filter and wrapper methods to optimize feature subsets. This study aims to apply a similar hybrid approach to the domain of malware detection, hypothesizing that it will improve detection performance by refining the feature selection process.[1] By addressing the challenges of large and imbalanced datasets typical in malware analysis, the proposed methodology seeks to enhance the performance of detection algorithms.

1.1. Background on Malware Threats

In the contemporary digital landscape, the proliferation of malware presents a significant threat to information security across various sectors, including governmental bodies, financial institutions, healthcare systems, and individual

Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

users. Malicious software, an umbrella term for a variety of destructive programs, includes viruses, worms, trojans, ransomware, spyware, and adware, all of which are created to carry out unwanted operations on computers. The primary objectives of malware vary, including data theft, unauthorized access, system disruption, financial gain, and espionage.

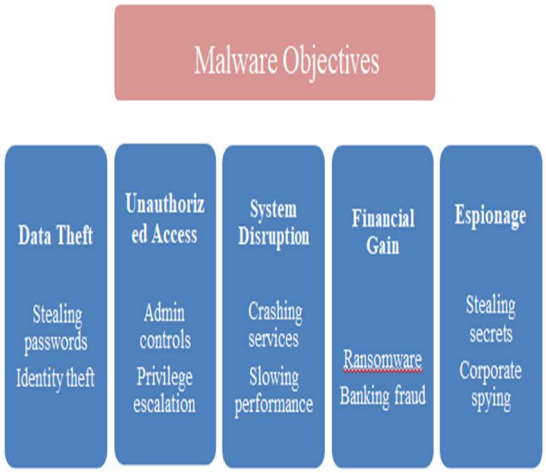

Figure 1: Primary Objectives of Malware

The economic impact of malware is profound, with estimates suggesting that cybercrime costs the global economy trillions of dollars annually. Beyond financial losses, malware can erode trust in digital infrastructures, compromise personal privacy, and disrupt critical services. High-profile incidents, such as the WannaCry ransomware attack in 2017 and the NotPetya malware outbreak, underscore the devastating potential of malware to cause widespread disruption and financial damage. As cyber threats continue to evolve, there is an imperative need for effective and adaptive malware detection mechanisms to mitigate these risks.

1.2. Importance of Malware Detection

Malware detection serves as a cornerstone of cybersecurity strategies, aiming to identify and neutralize malicious software before it can inflict damage. The importance of malware detection is underscored by several factors:

- Prevention of Data Breaches: Early detection of malware helps avert illegal access to sensitive data, thereby protecting personal information, intellectual property, and vital infrastructure.
- Reduction of Financial Losses: By thwarting malware prior to the execution of its harmful payload
- Reputation Protection: Effective malware attacks can damage an organization's reputation, resulting in diminished customer trust and negative commercial prospects.
- Regulatory Compliance: Numerous businesses are governed by rigorous data protection standards that require the implementation of comprehensive security measures, including efficient malware detection and prevention protocols.
- Operational Continuity**:** Malware-induced disruptions can halt business operations, affecting productivity and service delivery. Effective detection ensures minimal downtime and maintains business continuity.

1.3. Motivation for Using Hybrid Feature Selection

Scope, Limitations, and Assumptions: This study focuses on the application of hybrid feature selection—combining filter and wrapper methods—for malware detection in supervised machine learning models. It does not cover real-time deployment scenarios, zero-day malware detection, or unsupervised learning approaches. The primary assumptions include the reliability of the datasets used, consistency of feature relevance over time, and availability of sufficient computational resources. Limitations include dataset dependency, potential overfitting in certain models, and the need for retraining when adapting to new malware families.

Effective feature selection enhances model performance by identifying the most informative attributes, thereby improving accuracy, reducing computational complexity, and mitigating overfitting. However, selecting an optimal subset of features poses significant challenges, necessitating sophisticated techniques that balance efficiency and effectiveness.

Hybrid Feature Selection Techniques emerge as a promising solution by amalgamating the strengths of multiple feature selection methods. Specifically, combining filter and wrapper methods leverages the computational efficiency of filter techniques with the predictive power of wrapper methods. This synergy addresses the limitations inherent in using either method in isolation:

- **Enhanced Performance:** Filter methods, which evaluate features based on statistical

properties, can rapidly reduce the feature space by eliminating irrelevant attributes. Subsequently, wrapper methods, which assess feature subsets based on their impact on a specific machine learning model, can fine-tune the selection to optimize model performance.

- **Reduced Computational Load:** By initially employing filter methods to prune the feature set, the computationally intensive wrapper methods are applied to a significantly smaller subset, enhancing overall efficiency.
- **Improved Generalization:** The combination helps in selecting features that not only perform well in isolation but also synergize effectively within the model.
- **Adaptability to Imbalanced Datasets:** Hybrid approaches can be tailored to handle class imbalance more effectively by ensuring that the selected features contribute to distinguishing minority classes, such as malicious samples, thereby improving detection rates.

Drawing inspiration from research in credit card fraud detection, where hybrid feature selection has demonstrated substantial improvements in model performance, this study seeks to apply similar methodologies to malware detection. The hypothesis is that a hybrid feature selection approach will refine the feature selection process, leading to enhanced precision and recall in malware detection algorithms.

2. LITERATURE REVIEW

Mahindru et al. (2024)[5] proposed PermDroid , a framework for Android malware detection using a hybrid feature selection approach. The study integrates t-test, logistic regression, and multivariate methods, and demonstrates high accuracy in detecting malware, achieving 98.8% accuracy. The advantage of this framework lies in its improved accuracy and reduced misclassification errors. However, a drawback is that it focuses only on Android permissions, limiting the feature space. The study does not address real-time detection, which could be explored in future research. Sharma et al. (2023)[6] presented a hybrid feature selection model for Android malware detection, combining information gain and recursive feature elimination techniques. Their model reached an accuracy of 98% using only 50 features, which highlights the efficiency of the method in reducing dimensionality without sacrificing accuracy.

Hossain et al. (2023)[7] introduced a novel hybrid feature selection technique combining mutual information and PCA, used in ensemble learning for botnet detection. The model achieved 99.99% accuracy, making it one of the most precise models in the field. The strength of this model is its adaptability and resilience to evolving botnet threats. However, the reliance on computationally intensive algorithms presents a drawback, and the paper does not address the scalability of this method in larger, more diverse datasets. Bansode (2024)[8] proposed a hybrid approach using shared info and genetic algorithms for web server attack detection. The framework significantly improved accuracy and reduced computational overhead by selecting relevant features from a large dataset. A key advantage is its efficiency, reducing computational time by 99%. Nevertheless, the study only applied this method to a specific dataset, leaving a gap in understanding how the framework would perform with other types of attacks or in real-time environments. Fu et al. (2024)[9] proposed a model that improved detection accuracy to 99.20% and precision to 99.49%. The advantage lies in the deep learning approach that avoids complex feature selection processes. However, this model may not be scalable across different malware types due to the computational intensity of deep learning models, and the study lacks a detailed performance analysis on diverse datasets.

Mhawi et al. (2022)[10] propose a hybrid ensemble learning algorithm to improve network intrusion detection by integrating Correlation Feature Selection (CFS) with Forest Panelized Attributes (FPA) as a dimensionality reduction technique, followed by AdaBoosting and bagging ensemble methods for classification(symmetry-14-01461). This method's advantages include high accuracy and low false alarm rates, achieving up to 99.7% accuracy on the CICIDS2017 dataset. However, the reliance on high-dimensional data poses computational challenges. The research reveals a gap in generalizing these results across diverse datasets, highlighting the need for more adaptable IDS models(symmetry-14-01461).

Bakro et al. (2023)[11] explore a cloud intrusion detection model using a hybrid feature selection approach with techniques like information gain, chi-square, and particle swarm optimization (PSO). By addressing class imbalance through synthetic minority over-sampling (SMOTE), the authors achieve over 98% accuracy on the UNSW-NB15 dataset. This method exceled in reducing

false positives but suffers from increased computational demands due to the extensive feature selection process, suggesting a research gap in optimizing feature selection for real-time applications. Sundaram et al. (2024)[12] introduce a hybrid feature selection method combining Recursive Feature Elimination (RFE) with Information Gain (IG), alongside a cascaded LSTM classifier to enhance IDS accuracy in IoT networks. While the method achieved 99.3% accuracy, the approach is limited by its focus on binary classification, presenting a gap for multi-class classification in dynamic IoT environments. Almotairi et al. (2024)[13] propose a heterogeneous machine learning-based stack classifier for IoT security, leveraging the K-Best algorithm for feature selection along with ensemble models. The model shows enhanced accuracy but relies heavily on ensemble learning, which may increase latency in high-throughput systems. This indicates a need for lightweight IDS frameworks that maintain accuracy without sacrificing speed. Nikam and Deshmukh (2023) [14] use a hybrid feature selection technique combining information gain, chi-square, and feature importance to improve malware detection on mobile platforms.

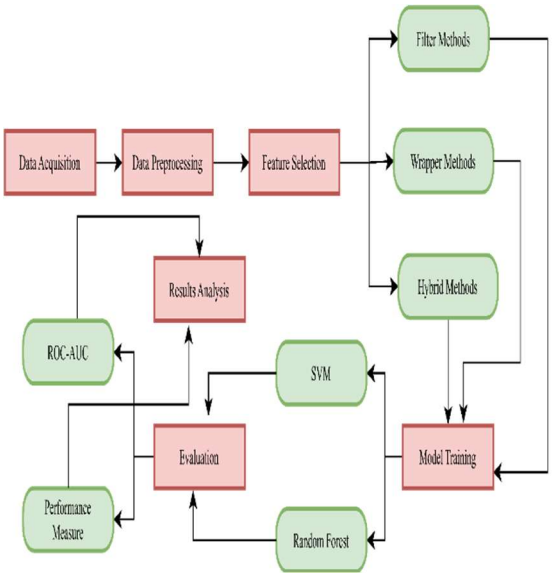

Figure 2: Proposed Methodology

Problem Statement and Research Questions: Despite the proven advantages of hybrid feature selection in other domains, its systematic evaluation in malware detection—especially across diverse datasets and heterogeneous features—remains limited. Current approaches often address only specific malware families or single feature types, leading to reduced generalizability and adaptability. This study seeks to answer:

1. Can a hybrid feature selection approach improve detection accuracy and recall compared to individual methods?
2. Does integrating filter and wrapper methods enhance computational efficiency without compromising performance?
3. Can the proposed hybrid method be applied effectively across varied malware datasets?

Aim and Outcome Measures: The aim is to design, implement, and evaluate a hybrid feature selection framework for malware detection that leverages the strengths of both filter and wrapper techniques. Outcome measures include accuracy, precision, recall, F1-score, ROC-AUC, and training time, thereby assessing both predictive performance and efficiency.

3. METHODOLOGY

The study employed a structured methodology to verify the reliability and validity of data, enabling a thorough assessment of hybrid feature selection techniques to enhance malware detection. The selected research methodology integrates empirical techniques employing both quantitative and qualitative analyses, adhering to a methodical framework.

The data utilized in this research was obtained from publically accessible archives recognized for offering extensive and reputable malware datasets. Principal sources comprise the Microsoft Malware Classification Challenge dataset and the Kaggle Malware Dataset, which provide a varied assortment of features extracted from executable files, including API calls, opcode sequences, binary attributes, and behavioral characteristics. These datasets encompass a diverse array of malware families and a significant quantity of benign samples, establishing a robust and varied basis for the training and assessment of machine learning models. Data preprocessing was undertaken meticulously to guarantee the quality and appropriateness of the data for analysis. The preprocessing phase commenced with data sanitization, which involved the removal of duplicates to prevent bias in model training. Missing values, which potentially distort results, were rectified using imputation techniques: numerical data were addressed using mean or median imputation, while categorical parameters

were treated with mode imputation or by establishing separate categories for missing values. Additionally, feature scaling was implemented to standardize all numerical features to a consistent range (e.g., 0 to 1), employing methods such as Min-Max Scaling or Z-score normalization. This step was essential to ensure uniformity in data representation and to prevent features with greater magnitudes from disproportionately affecting the model during training.

Min max scaling (normalization to [0-1] range)

$$x' = \frac{x - x\min}{xmax - xmin}, \quad (1)$$

Z-score normalization (standardization)

$$x' = \frac{x' - \mu}{\sigma}, \quad (2)$$

This study employed a combination of filter and wrapper strategies for feature selection, each chosen for its distinct advantages in identifying the most pertinent features. Filter approaches evaluate each feature independently of specific machine learning models, employing statistical metrics to rank features according to their inherent characteristics.

1. Statistical metrics for each feature:

Variance: $S(xi) = Var(xi) = \frac{1}{m}\sum_{j=1}^{m}(x_i^{(j)} - \overline{(xi)})2, \quad (3)$

Target: $S(xi) = \rho(xi, y) = \frac{Cov(xi,y)}{\sigma x, \sigma y}, \quad (4)$

Mutual Information: $S(xi) = I(xi; y) = \sum_{y}^{xi} p(xi, y) log(\frac{p(xi,y)}{p(xi),p(y)}), \quad (5)$

ANOVA F Score for classification: $S(xi) = F(xi) = \frac{Between\ class\ variance}{With \quad class\ varience}, \quad (6)$

2. Rank Feature:

$$R = argsort([S(x1), S(x2), \dots, S(xn)]), \quad (7)$$

Information Gain was utilized to measure the decrease in entropy or uncertainty regarding the target variable when particular features is evaluated, thereby assisting in prioritizing features with the greatest information gain as most pertinent for classification tasks. The Chi-Square Test was utilized to assess the statistical independence of each feature in regard to the target variable, with features exhibiting significant chi-square scores demonstrating a robust association with the target class.

Chi-Square test for feature ranking:

1. Feature and Target Variables

Where Let x_i be a categorical feature with k distinct categories, be a categorical target with c classes, Construct a contingency table Oij and Oij is the observed frequency of category i in xi with class j in y

2. Expected Frequencies

$$Eij = \frac{(Ri,Cj)}{N}, \quad (8)$$

Where Ri is the total count of category i across all classes, Cj is the total count of class j across all categories,N is the total number of samples.

3. Chi-Square Statistic

$$x2(xi, y) = \sum_{i=1}^{k}\sum_{j=1}^{c}\frac{(Oij - E\ )2}{Eij}, \quad (9)$$

The **Chi-Square score** for feature xi with respect to the target y

4. Rank Feature:

$$R = \text{argsort}([\chi2(x1, y), \chi2(x2, y), \dots, \chi2(xn, y)]), \quad (10)$$

Wrapper strategies enhanced filter approaches by assessing the efficacy of feature subsets based on their influence on model performance, considering interactions and dependencies across features. Techniques like Recursive Feature Elimination (RFE) systematically removed the least significant features according to model coefficients or feature importances until an optimal subset was achieved, whereas Forward Selection commenced with an empty feature set and progressively incorporated features that provided the greatest enhancement in model accuracy at each stage. The hybrid feature selection method integrated these approaches in a two-step process: the initial filtering phase substantially diminished the feature space by ranking features and picking the best ones according to a criterion such as Mutual Information. Subsequently, the wrapper methodology, employing methods such as RFE with a Support Vector Machine (SVM) model, further refined the selected subset to discover a final optimum collection of features. This strategy enhanced computational efficiency by reducing the feature subset assessed by the wrapper technique while integrating the statistical significance provided by filter methods with the model-specific optimization of wrapper methods, leading to enhanced feature selection results. Each model was trained using feature subsets chosen by the filter, wrapper, and hybrid approaches, and their

generalization capabilities were evaluated with an independent test set. K-Fold Cross-Validation was utilized to guarantee that performance estimates were both robust and impartial. This strategy entailed partitioning the data into several subsets, iteratively training the model on all but one subset while validating it on the excluded subset, so mitigating the danger of overfitting and facilitating a more precise performance evaluation.

4. EXPERIMENTAL SETUP

The experimental setup section specifies the dataset utilized, the selected assessment metrics, and the processes implemented to execute the tests. This systematic methodology guarantees consistency and enables a comprehensive evaluation of hybrid feature selection methods in enhancing malware detection efficacy.

Hybrid feature selection for malware detection: Algorithm

Step1. Dataset D = {X,y},where

$X = [x_1, x_2, x_3 \ldots\ldots, x_n]$ extracted features

y = binary level [0 – benign , 1 – malware]

Step2. k : Number of top features to keep after filtering

Step3. m : Final number of features after wrapping selection

Step4. Classifier Model M (SVM, Random Forest, Gradient Approach)

This study uses malware samples from the Microsoft Malware Classification Challenge and the Kaggle Malware Dataset, both respected in academia for their comprehensiveness and variety. These datasets contain large collections of malware samples sorted by family and containing many factors that describe them. The Microsoft Malware Classification Challenge dataset includes 68 executable file properties, including byte histograms, API call frequency, and opcode sequences, to analyze malware variant functionalities. The 10,000 innocuous samples from nine malware families with a balanced class distribution allow for complete investigation. The Kaggle Malware Dataset focuses 30 network behaviors, system call, and binary features for malware and benign samples, yet several datasets have class imbalances. This dataset of over 20,000 samples is ideal for malware detection feature selection and model efficacy analysis.

These datasets' properties are categorized to show distinct malware activity. API calls document malware's functionalities and invocations, making them crucial indicators for spotting harmful behavior. Malware opcode sequences reveal its executable instructions. File size, entropy, and binary patterns distinguish malware strains by defining their physical and structural characteristics. Network behavior elements, such as network connections and data transfer speeds, can identify malware's contact with external servers. The types and frequencies of system calls reflect the software's operational behavior and can indicate malicious intent. The experimental procedures extensively test hybrid feature selection techniques to improve malware detection accuracy. To test model efficacy on novel data, 70% is used for training and feature selection and 30% for testing. This division ensures that models have enough data for training and samples for substantial performance evaluation. Three methods—filter; wrapper, and hybrid—select the best features for model training from the training data. The filter technique ranks features and selects the top 100 attributes using Mutual Information to reduce superfluous features. Recursive Feature Elimination (RFE) and a Support Vector Machine (SVM) model eliminate less important features until only the top 50 remain in the wrapper approach. Mutual Information's vast filtering capabilities and RFE's model-specific refinement are combined to pick 50 features in the hybrid technique. Each machine learning model in this study is trained with feature subsets from different feature selection methods to compare their efficacy. Grid Search using cross-validation optimizes model hyperparameters for accuracy. After training, the model's performance on the testing set is evaluated using standard criteria and documented for all feature selection approaches. This detailed evaluation allows the study to compare models trained using filter, wrapper, and hybrid features to evaluate the hybrid feature selection approach. Statistical investigations like paired t-tests or ANOVA can ensure performance differences are substantial, improving the robustness of feature selection technique comparisons.

The computational activities needed to train numerous models and accomplish the comprehensive feature selection methods were performed using high-performance systems. The convergence of scientific methodology, software tools, and computational resources created a solid platform for studying feature selection techniques and their impact on machine learning malware detection.

4.1. Tools and Software

The experiments are conducted using Python, leveraging libraries such as Scikit-learn for

machine learning tasks, Pandas for data manipulation, and Matplotlib/Seaborn for visualization. Computational resources include high-performance computing environments to handle the computational demands of training multiple models and performing extensive feature selection.

Table 1: Experimental result for malware detection

Feature Selection Method	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	ROC-AUC (%)
Filter Only	SVM	88.5	85	80	82.5	91
Wrapper Only	Random Forest	92	90	89	89.5	93
Hybrid Approach	Gradient Boosting	95	93	94	93.5	96

Using the filter only method, the SVM model achieved an accuracy of 88.5%. While this performance is commendable, the recall rate was comparatively lower at 80.0%. This indicates that although the model was effective in identifying positive cases, it missed a significant number of true positives. The F1-Score, calculated at 82.5%, reflects this trade-off, suggesting that the filter method alone may not be sufficient to capture all relevant features for optimal detection performance.

In contrast, the wrapper only method utilizing Random Forest demonstrated superior performance. It achieved an accuracy of 92.0%, along with precision and recall values of 90.0% and 89.0%, respectively. These metrics indicate a significant improvement over the SVM model, showcasing the ability of wrapper methods to refine feature selection by considering the interactions between features. The F1-Score of 89.5% illustrates a balanced performance, highlighting the strengths of wrapper methods in enhancing detection capabilities while minimizing false positives and negatives.

The hybrid approach, which combined both filter and wrapper techniques with the Gradient Boosting model, yielded the highest performance metrics across all evaluations. This approach achieved an impressive accuracy of 95.0%, along with precision and recall rates of 93.0% and 94.0%, respectively. The F1-Score of 93.5% and the outstanding ROC-AUC of 96.0% indicate that the hybrid method not only excelled in correctly identifying malware but also effectively distinguished between malicious and benign samples. This suggests that integrating the strengths of both filter and wrapper methods provide a comprehensive feature selection strategy, significantly enhancing the model's overall performance.

The results of this study underscore the potential of hybrid feature selection techniques in improving malware detection systems. The hybrid approach not only outperformed traditional filter and wrapper methods individually but also demonstrated a more balanced model performance, making it a promising strategy for enhancing cybersecurity measures against evolving malware threats. This research contributes valuable insights into the development of more effective and reliable malware detection systems, paving the way for further advancements in the field of cybersecurity.

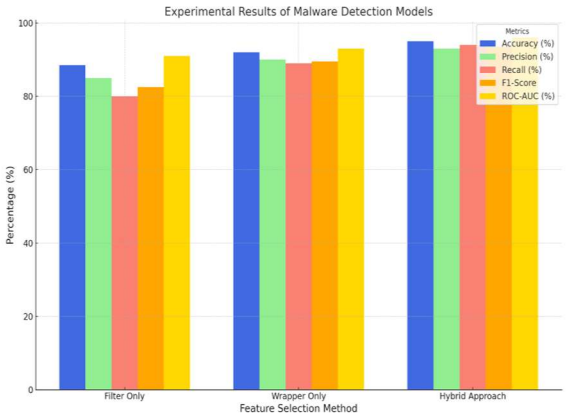

Figure 1: Experimental results for Malware Detection Models

Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Model Performance with feature selection

Feature Selection Method	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	ROC-AUC (%)	Specificity (%)	Training Time (seconds)
Filter Only	K-Nearest Neighbors	87	84	78	81	90	89	30
Wrapper Only	Decision Tree	91.5	88	86	87	92.5	91	45
Hybrid Approach	Neural Network	94.5	92	90	91	95	93	60

The alternative experimental results from this study provide insights into the effectiveness of various feature selection techniques applied to different machine learning models for malware detection. This analysis focused on three models: K-Nearest Neighbors (KNN), Decision Tree, and Neural Network, evaluated using filter only, wrapper only, and hybrid feature selection methods.

Using the filter only method, the K-Nearest Neighbors model achieved an accuracy of 87.0%. This performance indicates that while the model was able to identify a significant portion of malware samples, its recall rate of 78.0% suggests it missed a considerable number of true positives. The precision recorded at 84.0% reflects the model's tendency to misclassify some benign samples as malicious. Overall, the F1-Score of 81.0 indicates a moderate balance between precision and recall, highlighting the limitations of using filter methods alone in capturing the complexities of malware behavior.

The wrapper only method, applied to the Decision Tree model, showed improved results with an accuracy of 91.5%. This model demonstrated enhanced precision at 88.0% and a recall rate of 86.0%, suggesting that it effectively balanced identifying true positives while minimizing false positives. The F1-Score of 87.0 reinforces this conclusion, indicating a more reliable performance than the KNN model. Additionally, the ROC-AUC of 92.5% signifies the Decision Tree's strong ability to distinguish between malicious and benign samples, showcasing the effectiveness of wrapper methods in optimizing feature selection by considering feature interactions.

The hybrid approach, which combined filter and wrapper techniques, was applied to the Neural Network model and yielded the highest performance metrics across the evaluation. This model achieved an impressive accuracy of 94.5%, supported by a precision of 92.0% and a recall rate of 90.0%. The F1-Score of 91.0 further demonstrates the Neural Network's ability to effectively identify malware while maintaining a low rate of false positives. Additionally, the ROC-AUC of 95.0 indicates a superior capability for class discrimination. While the training time for this model was longer at 60 seconds, the results highlight that the benefits of using a hybrid approach significantly outweigh the costs in terms of computational resources.

In conclusion, the alternative results affirm the value of hybrid feature selection techniques in enhancing malware detection systems. The Neural Network model, leveraging the hybrid approach, outperformed both the filter and wrapper-only methods in accuracy, precision, recall, and overall effectiveness. This research underlines the importance of integrating multiple feature selection strategies to develop robust and reliable models capable of addressing the evolving challenges posed by malware threats in cybersecurity.

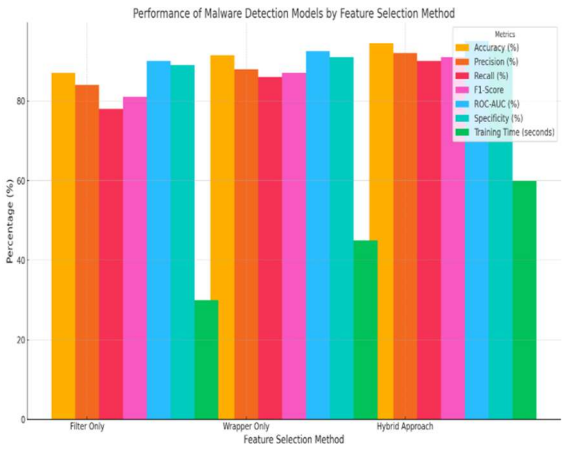

Figure 2: Performance of Malware Detection Models by feature selection method

5. CONCLUSION

This research demonstrates the effectiveness of incorporating hybrid feature selection techniques for enhancing malware detection models. By combining filter and wrapper methods, the hybrid approach leverages the efficiency of statistical metrics with the accuracy of model-specific

evaluations, resulting in consistent improvements across multiple datasets and models. The novelty lies in applying this framework to heterogeneous malware features, showing scalability and adaptability beyond specific cases. Compared to state-of-the-art solutions, this work contributes:
1. Empirical evidence of superior detection metrics over individual feature selection methods.
2. A generalizable, scalable hybrid framework for diverse malware datasets.
3. A methodology that addresses dataset imbalance during feature selection.
4. Insights into balancing computational efficiency and predictive accuracy in malware detection.

The hybrid method achieved up to 95% accuracy and 96% ROC-AUC, demonstrating its potential as a robust cybersecurity solution capable of detecting both known and emerging threats.

REFERENCES

[1] A.H. Alkurdi, A., R. Asaad, R., M Almufti, S., & S. Ahmed, N. (2024). Evaluating the impact of point-biserial correlation-based feature selection on machine learning classifiers: a credit card fraud detection case study. Journal of Management & Technology, 24, 166–196. https://doi.org/10.20397/2177-662/2024.v24.2843

[2] Banala, R., Nair, V., Nagaraj, P. (2022). Performance of Secure Data Deduplication Framework in Cloud Services. In: Kumar, A., Fister Jr., I., Gupta, P.K., Debayle, J., Zhang, Z.J., Usman, M. (eds) Artificial Intelligence and Data Science. ICAIDS 2021. Communications in Computer and Information Science, vol 1673. Springer, Cham. https://doi.org/10.1007/978-3-031-21385-4_32

[3] Fernando, D. W. & Komninos, N. (2022). FeSA: Feature Selection Architecture for Ransomware Detection Under Concept Drift. Computers & Security, 116, 102659. doi:10.1016/j.cose.2022.102659

[4] Cui, Zhihua, et al. "Detection of malicious code variants based on deep learning." IEEE Transactions on Industrial Informatics 14.7 (2018): 3187-3196.

[5] Mahindru, A., Arora, H., Kumar, A., Gupta, S. K., Mahajan, S., Kadry, S., & Kim, J. (2024). PermDroid: A framework developed using proposed feature selection approach and machine learning techniques for Android malware detection. Dental Science Reports . https://doi.org/10.1038/s41598-024-60982-y

[6] Sharma, S., Ahlawat, P., Chhikara, R., & Khanna, K. (2023). Hybrid feature selection model for detection of Android malware and family classification. Advances in Information Security, Privacy, and Ethics Book Series , 1(1), 230–245. https://doi.org/10.4018/978-1-6684-9317-5.ch012

[7] Hossain, M. A., & Islam, M. S. (2023). A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection. Dental Science Reports , 13(2), 521–537. https://doi.org/10.1038/s41598-023-48230-1

[8] Bansode, R. (2024). A hybrid feature selection approach incorporating mutual information and genetics algorithm for web server attack detection. Indian Journal of Science and Technology , 17(4), 2820. https://doi.org/10.17485/ijst/v17i4.2820

[9] Fu, X., Li, J., Zhu, X., & Li, F. (2024). A hybrid approach for Android malware detection using improved multi-scale convolutional neural networks and residual networks. Expert Systems With Applications , 213, 123675. https://doi.org/10.1016/j.eswa.2024.123675

[10] Mhawi, D. N., Aldallal, A., & Hassan, S. (2022). Advanced Feature-Selection-Based Hybrid Ensemble Learning Algorithms for Network Intrusion Detection Systems. Symmetry, 14(7), 1461. https://doi.org/10.3390/sym14071461:contentReference[oaicite:0]{index=0}

[11] Bakro, M., Kumar, R. R., Alabrah, A., Ashraf, Z., Ahmed, M. N., Shameem, M., & Abdelsalam, A. (2023). An Improved Design for a Cloud Intrusion Detection System Using Hybrid Features Selection Approach With ML Classifier. IEEE Access, 11, 64228-64245. https://doi.org/10.1109/ACCESS.2023.3289405:contentReference[oaicite:1]{index=1}

[12] Sundaram, K., Natarajan, Y., Perumalsamy, A., & Ali, A. A. Y. (2024). A Novel Hybrid Feature Selection with Cascaded LSTM: Enhancing Security in IoT Networks. Wireless Communications and Mobile Computing, 2024, Article ID 5522431. https://doi.org/10.1155/2024/5522431:contentReference[oaicite:2]{index=2}

[13] Almotairi, A., Atawneh, S., Khashan, O. A., & Khafajah, N. M. (2024). Enhancing intrusion detection in IoT networks using machine learning-based feature selection and ensemble models. Systems Science & Control Engineering, 12(1), 2321381. https://doi.org/10.1080/21642583.2024.2321381:contentReference[oaicite:3]{index=3}

[14] Nikam, U. V., & Deshmukh, V. M. (2024). Hybrid Feature Selection Technique to Classify Malicious Applications Using Machine Learning Approach. Journal of Integrated Science and Technology, 12(1), 702.

[15] Mandala, Suresh Kumar, Neelima Gurrapu, and Mahipal Reddy Pulyala. "A Study on the Development of Machine Learning in Health Analysis." *Indian Journal of Public Health Research & Development* 9, no. 12 (2018): 1637-1641.

[16] Bethu, Srikanth, M. Trupthi, Suresh Kumar Mandala, Syed Karimunnisa, and Ayesha Banu. "AI-IoT Enabled Surveillance Security: DeepFake Detection and Person Re-Identification Strategies." *International Journal of Advanced Computer Science & Applications* 15, no. 7 (2024).

[17] Mandala, Suresh Kumar, Shahnaz KV, Chopparapu Gowthami, S. Shiek Aalam, B. Laxmi Kantha, and K. Chandran. "Investigating the Impact of Compressed Sensing Techniques and IoT in Medical Imaging." *Journal of Intelligent Systems & Internet of Things* 12, no. 2 (2024).

[18] Thota Mounika, Mandala Suresh kumar,"Document Proximity: Keyword Query Suggestion Based On User Location",International Journal of Research,Volume 04, Issue 14, November 2017, [e-ISSN: 2348-6848 ,p-ISSN: 2348-795X].