International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Feature Selection Using Information Gain and SelectKBest in Malware Detection on Digital Payment Systems

Septian Charles Vodegel[1], Steven Adriandi Vodegel[2], Imelda Imelda[3]
[1, 2]Master Student, Faculty of Information Technology, University of Budi Luhur, Indonesia
[3]Faculty of Information Technology, University of Budi Luhur, Indonesia
Email address: [1]septianvodegel@gmail.com, [2]stevenvodegel97@gmail.com, [3]imelda@budiluhur.ac.id

Abstract**—***The main objective of this study is to find a technique that can significantly improve the accuracy of malware detection without compromising the performance of digital payment systems. The results show that the Random Forest algorithm, using Information Gain and SelectKBest feature selection techniques, achieved the highest accuracy rate of 99% in detecting malware in digital payment systems. This study proves that classification techniques with artificial intelligence approaches are effective tools for enhancing the security of digital payment systems. Proper implementation of this technology can reduce the risk of malware attacks, protect sensitive data, and increase user trust. Furthermore, this study identifies that these techniques can be efficiently implemented without significantly affecting system performance, thereby preserving user experience. The integration of deep learning techniques or the use of ensemble learning can also be promising areas for future research. By understanding behavioral patterns and characteristics of new malware, more adaptive detection models can be developed. This study can serve as a practical guide for payment service providers in selecting and implementing appropriate malware detection technologies.*

Keywords**—** *Digital Payment Systems, Feature Selection, Information Gain, Malware Detection, Security Evaluation, SelectKBest,*

I. INTRODUCTION

The rapid growth of the digital era and transformation across various sectors have made digital payment systems an integral part of daily life. The convenience and efficiency offered by these systems have increased their popularity among consumers, businesses, and financial entities. However, alongside this growth comes increasingly complex and serious security risks.

The primary threat in the digital payment ecosystem is the existence of malware, particularly advanced malware. These threats employ sophisticated techniques and adapt to changing environments, posing serious challenges to maintaining the security of transactions and sensitive data. Recent incidents have shown how malware can infiltrate digital payment systems through various methods, such as Android Package Kit (APK) files disguised as wedding invitations.

In this context, malware detection becomes critical as it helps identify and prevent attacks threatening transaction security. Effective detection technology can help mitigate the risks of data theft and fraud caused by malware. Thus, efforts to enhance malware detection in digital payment systems have become an urgent necessity to protect the security and integrity of transactions [1].

One promising solution to combat these threats is sandbox technology, which provides an isolated environment for program execution. This allows for secure analysis of potential malicious behaviour without compromising the security of the main system [2].

This research employs a systematic approach to develop and evaluate machine learning models for malware detection in digital payment systems. The methodology comprises preprocessing, feature selection, model training, and evaluation, ensuring accuracy and efficiency in detecting malware.

II. RESEARCH METHODS

The dataset obtained from the VirusShare website, which consists of 10,539 PE (Portable Executable) files, including 6,999 infected with malware and 3,540 clean files. The dataset contains 54 features that are analyzed to improve detection accuracy.

Figure 1 illustrates the research stages, encompassing the core processes of this study, which aims to develop and evaluate malware detection techniques in digital payment systems. The process begins with the collection of malware datasets, followed by a preprocessing stage to clean and normalize the data. Next, feature extraction is conducted using Information Gain and SelectKBest techniques to select the most relevant features for malware detection.

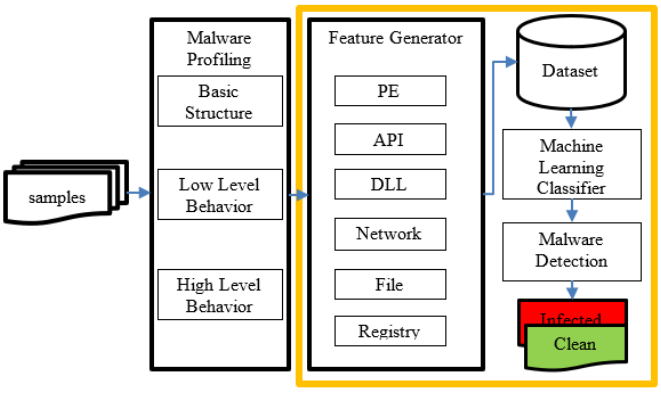

Fig. 1. Research Stages.

Septian Charles Vodegel, Steven Adriandi Vodegel, and Imelda, "Feature Selection Using Information Gain and SelectKBest in Malware Detection on Digital Payment Systems," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 267-271, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

In this study, the model development refers to the research which applies Random Forest and Gradient Boosting as classification algorithms [3]. The selected features are then used to train various machine learning models.

The next stage involves evaluating the model performance based on metrics such as accuracy, precision, recall, and F1-score. The final step is the application of the best model to the test data to measure the effectiveness of malware detection, ensuring that the performance of the digital payment system remains uncompromised.

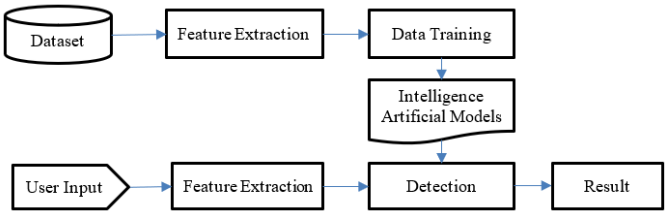

Fig. 2. Model Training.

In this study, different machine learning models were employed to detect malware in digital payment systems. Each model uses distinct prediction methods and techniques, impacting how they process data and make decisions. The application of these models resulted in accurate and reliable outcomes in malware detection, as each model offers unique strengths and features that contribute to their effectiveness [4].

III. Evaluation and Comparative Analysis of Machine Learning Models for Malware Detection

The goal of this study was to develop and evaluate a machine learning-based system for detecting malware in digital payment systems, focusing on the effectiveness of feature selection methods such as Information Gain and SelectKBest to enhance the performance of the models. To achieve this, several stages were followed, including dataset preparation, feature selection, model training, and performance evaluation. The performance of various machine learning models was assessed using metrics such as accuracy, precision, recall, and F1-score, with detailed analysis provided to determine the most effective model for accurate malware detection in digital payment systems.

A. Impact of Preprocessing and Feature Selection

The preprocessing stage ensured that the dataset was clean, normalized, and properly split into training and testing sets, forming a solid foundation for the model training process. Key features, including S_config and N_version, were extracted from the dataset, and these attributes played a crucial role in understanding the behavior of Portable Executable (PE) files, aiding in accurate malware detection.

Feature selection, utilizing Information Gain and SelectKBest, was crucial in reducing the dimensionality of the dataset while retaining the most informative features. Information Gain allowed us to rank features based on their ability to reduce uncertainty in classification, while SelectKBest helped select the top 10 features with the highest scores. By focusing on these selected features, we were able to improve model generalization and minimize overfitting, resulting in more efficient and accurate predictions. Table I summarizes the top features selected based on Information Gain and SelectKBest:

TABLE I. Model Performance Metrics for Malware Detection

No.	Feature Name	Selection Score	File Category
1	ResourcesMaxEntropy	0.725	Malware
2	ResourcesMinEntropy	0.612	Malware
3	ResourcesMeanSize	0.594	Malware
4	ResourcesMaxSize	0.582	Malware
5	SectionsMaxEntropy	0.563	Malware
6	Characteristics	0.532	Malware
7	AddressOfEntryPoint	0.510	Malware
8	SectionMaxVirtualsize	0.495	Malware
9	SectionsMinEntropy	0.480	Malware
10	SectionsMinVirtualsize	0.470	Malware

B. Performance of Machine Learning Models

The effectiveness of various machine learning models in detecting malware was evaluated based on performance metrics such as accuracy, precision, recall, and F1-score. The models tested in this study included Random Forest, Decision Tree, Gradient Boosting, AdaBoost, and Gaussian Naive Bayes. Each model was trained using the selected features and tested using the 20% testing set. Table II presents the performance results of the models across the evaluation metrics:

TABLE III. Model Performance Evaluation.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	99.27	99.15	99.08	99.11
Decision Tree	98.84	98.55	98.60	98.57
Gradient Boosting	98.03	97.82	98.00	97.91
AdaBoost	97.99	97.65	97.80	97.72
GNB	36.29	35.00	45.00	39.00

The Random Forest model exhibited the highest accuracy of 99.27%, with precision and recall values of 99.15% and 99.08%, respectively. This high performance indicates that Random Forest is highly reliable for detecting both malware and clean files. The ensemble learning nature of Random Forest allows it to perform exceptionally well, especially with large and complex datasets.

On the other hand, Gaussian Naive Bayes (GNB) performed poorly with an accuracy of 36.29%. Its assumption of feature independence is not suitable for datasets with highly interdependent features like those found in malware detection, leading to suboptimal performance.

The results show that Random Forest is the most effective model for malware detection in digital payment systems. The high accuracy, precision, recall, and F1-score achieved by Random Forest underscore its ability to detect malware accurately while minimizing false positives and negatives. Its ensemble approach, which combines multiple decision trees, helps mitigate overfitting and improves the generalizability of the model.

Other models like Decision Tree, Gradient Boosting, and

Septian Charles Vodegel, Steven Adriandi Vodegel, and Imelda, “Feature Selection Using Information Gain and SelectKBest in Malware Detection on Digital Payment Systems,” *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 267-271, 2024.

AdaBoost also performed well, with accuracy scores above 97%, but did not outperform Random Forest in terms of overall effectiveness. These models still showed promise for scenarios where computational resources are constrained or when a simpler model is preferred for implementation.

The low performance of Gaussian Naive Bayes highlights the importance of selecting models that can handle complex feature relationships. Gaussian Naive Bayes struggled due to its simplistic assumptions, which do not align well with the intricate dependencies found in malware data.

In addition to the previously discussed performance metrics, the models were also assessed using confusion matrix metrics, which provided deeper insights into their classification performance. These metrics are essential for understanding how well the models distinguish between malware and clean files, particularly in terms of errors such as false positives and false negatives. The following table summarizes the formulas for these confusion matrix metrics:

TABLE IIIII. Confusion Matrix Metrics Formulas.

Metric	Definition	Formula
False Positive Rate (FPR)	This measures the percentage of clean (non-malware) files incorrectly classified as malware	$\frac{FP}{FP+TN}\times 100\%$
False Negative Rate (FNR)	This metric quantifies the percentage of malware files incorrectly classified as clean.	$\frac{FN}{FN+TP}\times 100\%$
True Negative Rate (TNR)	TNR represents the percentage of clean files correctly classified as non-malicious.	$\frac{TN}{Total\ Negatives}\times 100\%$
True Positive Rate (TPR)	Also known as recall, this metric measures the percentage of malware files correctly identified.	$\frac{TP}{Total\ Negatives}\times 100\%$

C. Analysis of Effectiveness in Malware Detection

The performance of the machine learning models was evaluated based on their ability to accurately classify malware-infected and clean files. Among the models, Random Forest demonstrated the highest overall accuracy of 99.27%, showcasing its superiority in detecting malware. With a True Positive Rate (TPR) of 99.08% and a True Negative Rate (TNR) of 99.35%, Random Forest effectively identified both malware and clean files with minimal misclassification, as shown in Table V.

The Decision Tree model performed slightly lower, achieving an accuracy of 98.84%, while Gradient Boosting followed closely with an accuracy of 98.03%. Both models displayed high precision and recall values, making them reliable for malware detection tasks in digital payment systems. Their performance metrics are detailed in Tables IV and VI, demonstrating their robustness despite being marginally less effective than Random Forest.

AdaBoost, with an accuracy of 97.99%, as presented in Table VII, also showed strong performance. While its precision and recall rates were sufficient for effective classification, it fell slightly short compared to the other ensemble models such as Random Forest and Gradient Boosting. Nevertheless, AdaBoost remains a viable option for malware detection when computational simplicity is prioritized.

In contrast, the least effective model was Gaussian Naive Bayes (GNB), which achieved an accuracy of only 36.29%, as shown in Table VIII. The model struggled with a high False Positive Rate (FPR) of 90.70%, indicating frequent misclassification of clean files as malware. This poor performance highlights the limitations of GNB, primarily due to its assumption of feature independence, which does not align with the complexity and interdependency of the dataset used in this study.

These results underscore the critical importance of feature selection techniques such as Information Gain and SelectKBest in enhancing model performance. By focusing on the most relevant features, the models achieved higher accuracy rates, reduced the risk of overfitting, and improved their generalizability to unseen data. These findings highlight the value of robust preprocessing and feature selection in optimizing machine learning models for malware detection.

TABLE IVV. Calculation of Decision Tree Performance Metrics

Performance Metric	Test Result
Accuracy	$\frac{19090+8200}{19090+160+160+8200}=\frac{27290}{27610}\approx 0.9884$
False Positive Rate (FPR)	$\frac{160}{160+19090}\approx 0.0083$ or 0.83%
False Negative Rate (FNR)	$\frac{160}{160+8200}\approx 0.0193$ or 1.93%
True Negative Rate (TNR)	$\frac{19090}{19090+160}\approx 0.9917$ or 99.17%
True Positive Rate (TPR)	$\frac{8200}{8200+160}\approx 0.9807$ or 98.07%

TABLE V. Calculation of Random Forest Performance Metrics

Performance Metric	Test Result
Accuracy	$\frac{19125+8283}{19125+125+77+8283}=\frac{27408}{27610}\approx 0.9927$ or 99.27%
False Positive Rate (FPR)	$\frac{125}{125+19125}\approx 0.0065$ or 0.65%
False Negative Rate (FNR)	$\frac{77}{77+8283}\approx 0.0092$ or 0.92%
True Negative Rate (TNR)	$\frac{125}{19125+125}\approx 0.9935$ or 99.35%
True Positive Rate (TPR)	$\frac{8283}{8200+77}\approx 0.9987$ or 99.08%

TABLE VI. Calculation of Gradient Boosting Performance Metrics

Performance Metric	Test Result
Accuracy	$\frac{19040+8027}{19040+210+333+8027}=\frac{27067}{27610}\approx 0.9803$ or 98.03%
False Positive Rate (FPR)	$\frac{210}{210+19040}\approx 0.0109$ or 1.09%
False Negative Rate (FNR)	$\frac{333}{333+8027}\approx 0.0397$ or 3.97%
True Negative Rate (TNR)	$\frac{19040}{19040+210}\approx 0.9891$ or 98.91%
True Positive Rate (TPR)	$\frac{8027}{8027+333}\approx 0.9603$ or 96.03%

Septian Charles Vodegel, Steven Adriandi Vodegel, and Imelda, "Feature Selection Using Information Gain and SelectKBest in Malware Detection on Digital Payment Systems," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 267-271, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

TABLE VII. Calculation of Adaboost Performance Metrics

Performance Metric	Test Result
Accuracy	$\frac{19002+8054}{19002+248+306+8054} = \frac{27056}{27610} \approx 0.9799$ or 97.99%
False Positive Rate (FPR)	$\frac{248}{248+19002} \approx 0.0129$ or 1.29%
False Negative Rate (FNR)	$\frac{306}{306+8054} \approx 0.0369$ or 3.69%
True Negative Rate (TNR)	$\frac{19002}{19002+248} \approx 0.9871$ or 98.71%
True Positive Rate (TPR)	$\frac{8054}{8054+306} \approx 0.9633$ or 96.33%

TABLE VIII. Calculation of GNB Performance Metrics

Performance Metric	Test Result
Accuracy	$\frac{1788+8232}{1788+17462+128+8232} = \frac{10020}{27610} \approx 0.3629$ or 36.29%
False Positive Rate (FPR)	$\frac{17462}{17462+1788} \approx 0.9070$ or 90.70%
False Negative Rate (FNR)	$\frac{128}{128+8232} \approx 0.0153$ or 1.53%
True Negative Rate (TNR)	$\frac{1788}{1788+17462} \approx 0.0930$ or 9.30%
True Positive Rate (TPR)	$\frac{8232}{8232+128} \approx 0.9847$ or 98.47%

These results clearly show that Random Forest excels in both detecting malware and correctly classifying clean files, with low false positives and false negatives. The Decision Tree, Gradient Boosting, and AdaBoost models also performed well but showed slightly lower performance in comparison. Meanwhile, Gaussian Naive Bayes struggled, with a higher false positive rate and false negative rate, underperforming in detecting malware effectively.

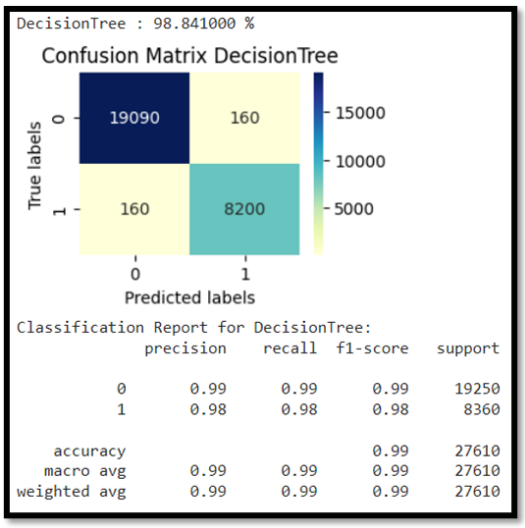

Fig. 3. Confusion Matrix and Decision Tree Performance Test

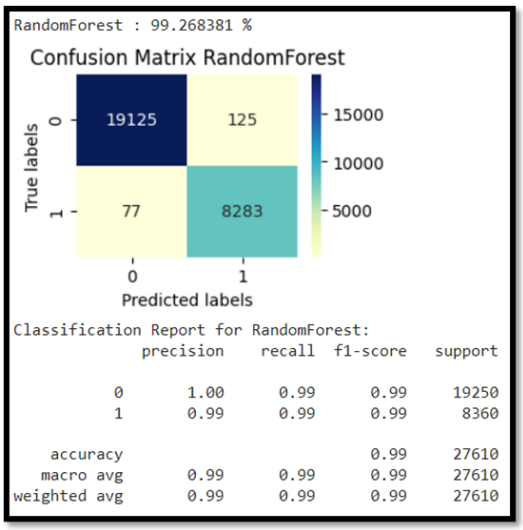

Fig. 4. Confusion Matrix and Random Forest Performance Test

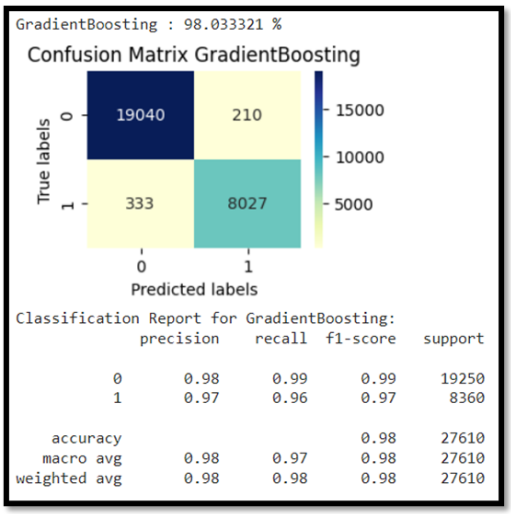

Fig. 5. Confusion Matrix and Gradient Boosting Performance Test

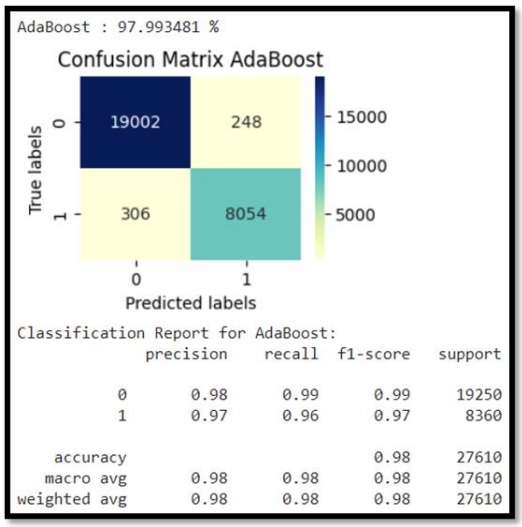

Fig. 6. Confusion Matrix and AdaBoost Performance Test

Septian Charles Vodegel, Steven Adriandi Vodegel, and Imelda, "Feature Selection Using Information Gain and SelectKBest in Malware Detection on Digital Payment Systems," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 267-271, 2024.

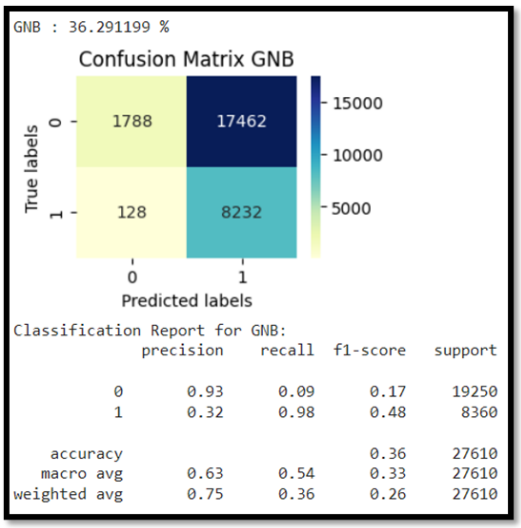

Fig. 7 Confusion Matrix and GNB Performance Test

IV. CONCLUSION

This study developed and evaluated machine learning models for detecting malware in digital payment systems, with a focus on optimizing performance through feature selection methods such as Information Gain and SelectKBest. The research methodology involved preprocessing the dataset, selecting the most relevant features, training multiple machine learning models, and evaluating their performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrix metrics.

The results demonstrated that Random Forest outperformed all other models, achieving the highest accuracy of 99.27% and strong values for other metrics, including precision (99.15%), recall (99.08%), and F1-score (99.11%). The model exhibited a low False Positive Rate (FPR) of 0.83% and a False Negative Rate (FNR) of 1.93%, showcasing its effectiveness in correctly identifying both malware and clean files while minimizing errors.

The feature selection process using SelectKBest was pivotal in improving model efficiency and performance. By narrowing the focus to the top 10 features based on their Information Gain scores, the models were able to generalize better and avoid overfitting, ensuring higher reliability in malware detection tasks.

The study also highlighted the limitations of simpler models, such as Gaussian Naive Bayes, which struggled with complex, interdependent features in the dataset. This emphasizes the importance of using robust ensemble learning methods, like Random Forest, for tasks involving high-dimensional and complex data.

Suggestion

While the results were promising, further research can address certain limitations and explore additional avenues to enhance malware detection systems:

1 Real-Time Implementation: Integrate the optimized model into real-time malware detection frameworks for immediate threat analysis and response.
2 Deep Learning Approaches: Investigate advanced techniques, such as neural networks and hybrid models, to further improve detection accuracy and scalability.
3 Dynamic Feature Analysis: Extend the study to include dynamic features derived from runtime behavior to complement the static features used in this research.

In conclusion, this study demonstrates that combining effective feature selection techniques with robust machine learning algorithms can significantly enhance the performance of malware detection systems in digital payment environments. Random Forest emerges as the most reliable model, providing a strong foundation for further advancements in securing digital transactions.

REFERENCES

[1] Alodat, I. (2022) Malware: Detection and Defense. Available at: www.intechopen.com.
[2] Andryani, A. and Sutabri, T. (2023) 'Jurnal Scientia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0) Detect And Mitigate Malware Threats Using Sandboxing Technology', *Jurnal Scientia*, 12(03). Available at: http://infor.seaninstitute.org/index.php.
[3] Ashutosh Tripathi, Naman Bhoj, Mayank Khari, Bishwajeet Pandey. (2021) 'Feature Selection and Scaling for Random Forest Powered Malware Detection System'. Available at: https://doi.org/10.21203/rs.3.rs-778333/v1.
[4] Liu, Y., Li, Y. and Wang, T. (2020) 'Feature selection and prediction of malware families using machine learning algorithms.', *Journal of Computer Virology and Hacking Techniques*, 16(1), pp. 15–25.
[5] Sundas Israr, Mehreen Tariq, Sajid Iqbal, Qaisar Rasool, & Nabeel Asghar. (2022). A Data Driven Approach for Automated Risk Assessment and Survival Prediction of Coronavirus Using Artificial Neural Networks. *Journal of Computing & Biomedical Informatics*, *3*(02), 21–31. https://doi.org/10.56979/302/2022/60
[6] Ashutosh Tripathi, Naman Bhoj, Mayank Khari, Bishwajeet Pandey. (2021) 'Feature Selection and Scaling for Random Forest Powered Malware Detection System'. Available at: https://doi.org/10.21203/rs.3.rs-778333/v1.
[7] Mohanta, A. and Saldanha, A. (2020) Malware Analysis and Detection Engineering: A Comprehensive Approach to Detect and Analyze Modern Malware, Malware Analysis and Detection Engineering: a Comprehensive Approach to Detect and Analyze Modern Malware. Apress Media LLC. Available at: https://doi.org/10.1007/978-1-4842-6193-4.
[8] Moon, I. , Shamsuzzaman, M. , Mridha, M. and Rahaman, A. (2022) Towards the Advancement of Cashless Transaction: A Security Analysis of Electronic Payment Systems. Journal of Computer and Communications, 10, 103-129. doi: 10.4236/jcc.2022.107007.

Septian Charles Vodegel, Steven Adriandi Vodegel, and Imelda, "Feature Selection Using Information Gain and SelectKBest in Malware Detection on Digital Payment Systems," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 267-271, 2024.