International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
***Article Received:** 10 July 2023 **Revised:** 31 August 2023 **Accepted:** 18 September 2023*

HMCMA: Design of an Efficient Model with Hybrid Machine Learning in Cyber security for Enhanced Detection of Malicious Activities

Mr. Mahesh T. Dhande [1]* Dr. Sanjaykumar Tiwari[2] Dr. Nikhil J. Rathod [3]
[1] Ph.D. Research Scholar, Department of Computer Science & Engg, Monad University, Hapur (U.P.) and Assistant Professor, Matoshri College of Engg. & Research Centre, Nashik
maheshdhande88@gmail.com
[2] Associate Professor, Department of Computer Science & Engg, Monad University, Hapur (U.P.) skt008@gmail.com
[3] Associate Professor, Department of Mechanical Science & Engg, Monad University, Hapur (U.P.)
rathod.nikhil358@gmail.com

Abstract: In the rapidly evolving landscape of cyber security, the incessant advancement of malicious activities presents a formidable challenge, necessitating a paradigm shift in detection methodologies. Traditional methods, primarily reliant on static rule-based systems, exhibit palpable limitations in grappling with the dynamic and sophisticated nature of modern cyber threats. This inadequacy underscores the urgent need for innovative approaches that can adeptly adapt and respond to the ever-changing threat environment. Addressing this exigency, the present research introduces a novel hybrid machine learning model, ingeniously crafted to transcend the constraints of existing malicious activity detection frameworks. The proposed model synergizes the strengths of diverse machine learning strategies, including anomaly detection techniques including Isolation Forest and One-Class SVM, and validates the results of these classifiers using Random Forest and Gradient Boosting operations. The validated malware instances are classified into malware types using fusion of Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) based Recurrent Neural Networks (RNNs) under real-time network configuration sets. This eclectic amalgamation not only leverages the unique capabilities of each algorithm but also harmonizes them to forge a more robust and precise detection mechanisms. The strategic integration of these algorithms facilitates a comprehensive analysis of network traffic and system logs, thereby significantly enhancing the detection accuracy. Furthermore, the model's adaptive learning component ensures its relevance and efficacy in the face of evolving cyber threats, a quintessential feature for contemporary cyber security solutions. Empirical evaluations, conducted using multiple malware datasets and samples, substantiate the model's superiority over existing methods. It exhibited a remarkable 10.4% improvement in precision, an 8.5% increase in accuracy, a 4.9% enhancement in recall, an 8.3% rise in AUC, a 4.5% boost in specificity, and a notable 2.5% reduction in detection delay. These compelling results underscore the model's potential in revolutionizing malicious activity detection, providing organizations with a more effective and resilient defense mechanism against a spectrum of cyber threats. The research culminates in a significant stride forward in cyber security, offering a robust, adaptive, and comprehensive solution that addresses the pressing need for advanced malicious activity detection, thereby bolstering the overall cyber security posture of organizations in the digital age sets.

Keywords: Machine Learning, Cyber security, Anomaly Detection, Deep Learning, Adaptive Algorithms

1. Introduction

In the intricate and ever-evolving domain of cyber security, the prevalence of sophisticated malicious activities has escalated exponentially, posing severe threats to the digital infrastructure of organizations worldwide. This surge in cyber threats underscores a critical challenge for traditional security mechanisms, which are increasingly found to be inadequate in the face of novel and complex cyber-attacks. Consequently, there emerges an imperative need for advanced detection methodologies that are not only efficient but also adaptive to the continuously evolving cyber threat landscape.

Historically, cyber security approaches have predominantly relied on static rule-based systems, characterized by their stringent parameters and predefined threat patterns. While effective against known threats, these systems exhibit a significant shortfall in identifying new, sophisticated attacks, primarily due to their lack of adaptability and reliance on prior knowledge of threat signatures. This gap in detection capabilities has been a catalyst for an urgent call to action within the cyber security research community, prompting a shift towards more dynamic and intelligent solutions.

Enter the realm of machine learning – a field that has shown immense potential in transforming the cyber security landscape. Machine learning, with its inherent capability to learn and adapt from data, offers a promising avenue for developing advanced threat detection models. These models can learn from patterns and anomalies in data, enabling them to identify malicious activities that deviate from normal behavior.

However, the application of machine learning in cyber security is not without its challenges. One primary concern is the selection of appropriate algorithms that can efficiently process the vast and varied data inherent in network environments.

Recognizing these challenges, the current research proposes a novel approach – a hybrid machine learning model specifically designed for malicious activity detection in cyber security. This model is not a mere assemblage of various machine learning techniques; rather, it is a meticulously crafted system that integrates the strengths of both anomaly detection and deep learning algorithms. The inclusion of anomaly detection methods, such as Isolation Forest and One-Class SVM, enables the model to effectively identify outliers in network traffic and system logs. Concurrently, the incorporation of deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) empowers the system to unravel complex patterns in the data, a task at which traditional machine learning algorithms might falter.

Moreover, the hybrid model also leverages the robustness of traditional machine learning algorithms, such as Random Forest and Gradient Boosting. These algorithms contribute to the model's ensemble learning capability, enhancing its overall predictive accuracy. The harmonious integration of these diverse algorithms results in a model that is not only more accurate but also more adaptable to the dynamic nature of cyber threats.

The introduction of this hybrid machine learning model marks a significant advancement in the field of cyber security. By addressing the limitations of existing detection methods and incorporating the strengths of various machine learning paradigms, this research paves the way for a new era in cyber threat detection. It promises a more robust, accurate, and adaptive mechanism for identifying and mitigating malicious activities, thereby fortifying the cyber security defenses of organizations in our increasingly digital world scenarios.

Motivation & Contributions

The impetus behind this research is rooted in the pressing need to counteract the burgeoning sophistication of cyber threats that continuously besiege the digital ecosystem. Traditional cyber security methods, although foundational in their era, now grapple with the velocity and intricacy of modern cyber-attacks. This disparity between evolving threats and stagnant detection methodologies not only exposes vulnerabilities in existing security systems but also highlights a critical juncture in cyber security research. It is this juncture that motivates the present study, driving the pursuit of a more resilient and intelligent approach to malicious activity detection.

At the heart of this research lies the recognition that the landscape of cyber threats is not static; it is a dynamic, ever-changing arena where adversaries continually adapt and evolve. This realization necessitates a paradigm shift from conventional, rule-based detection systems to more agile, learning-driven approaches. The proposed hybrid machine learning model is conceived as a response to this need, representing a significant leap in the field of cyber security.

The contributions of this research are manifold and impactful. Firstly, the study introduces an innovative hybrid machine learning model, ingeniously designed to integrate the strengths of various machine learning paradigms. This model is a testament to the potential of machine learning in enhancing cyber security defenses, offering a more nuanced and sophisticated approach to threat detection than traditional methods.

Secondly, the research addresses a crucial gap in the current cyber security literature – the lack of a comprehensive and adaptive model that can effectively keep pace with the evolving nature of cyber threats. By leveraging a combination of anomaly detection, deep learning, and traditional machine learning algorithms, the proposed model not only enhances detection accuracy but also provides the flexibility to adapt to new threats.

Thirdly, the empirical evaluation of the model, using diverse malware datasets and samples, demonstrates its superiority over existing methods. The observed improvements in precision, accuracy, recall, AUC, specificity, and detection delay are not just statistical victories; they represent a tangible advancement in the capability to protect digital assets from malicious activities.

Finally, this research contributes to the broader understanding of how hybrid machine learning models can be effectively employed in real-world cyber security scenarios. It offers valuable insights into the design and implementation of adaptive learning systems, paving the way for future innovations in the field.

In sum, the motivation behind this research is driven by the urgent need to address the limitations of current cyber security methods. The contribution of this study lies in its development of a hybrid machine learning model that embodies the adaptability, accuracy, and robustness required to combat the sophisticated cyber threats of the modern world, thereby significantly enhancing the cyber security posture of organizations.

2. Deep Dive into Malware Detection Models

The landscape of cyber security is an ever-shifting battleground, with new forms of malicious activities emerging at a pace that traditional detection methods struggle to match. This challenge has spurred extensive research into advanced detection techniques, particularly leveraging machine learning algorithms. A review of recent literature reveals various approaches and innovations in this realm, each contributing uniquely to the field sets.

Zhang et al. [1] introduced a lightweight malware traffic classification method based on broad learning architecture, addressing the need for efficient processing in IoT environments. Their approach underscores the importance of lightweight, yet effective solutions in the ever-expanding Internet of Things. Li et al. [2] focused on the challenge of imbalanced malware family classification, employing multimodal fusion and weight self-learning to improve classification accuracy. This work is pivotal in addressing the skewness often found in real-world datasets.

Kural et al. [3] presented an innovative audio-based malware family detection framework, demonstrating the potential of unconventional feature extraction methods in malware classification. Similarly, Barut et al. [4] explored the use of attention-based neural networks for malware traffic classification, prioritizing privacy-preserving techniques [4]. These studies indicate a growing trend towards utilizing advanced neural network architectures and feature extraction methods for enhanced detection capabilities.

Yan et al. [5] provided a comprehensive survey of adversarial attack and defense methods in malware classification, highlighting the cat-and-mouse game between cyber-attackers and defenders. The work of Zhong et al. [6], which illuminated malware byte codes with images for classification, further exemplifies the innovative approaches being explored in the field [6]. This visual representation of malware offers a novel perspective on malware detection and classification.

The use of cross-modal CNNs by Kim et al. [7] for malware classification using non-disassembled files marks a significant stride in simplifying the classification process, bypassing the complex and time-consuming step of file disassembly [7]. Belal and Sundaram's development of the Global-Local Attention-Based Butterfly Vision Transformer [8] and He et al.'s ResNeXt+ model [9] both emphasize the growing importance of attention mechanisms in enhancing the accuracy of malware detection systems.

Ravi et al.'s attention-based multidimensional deep learning approach [10] specifically targets IoMT malware detection in healthcare systems, addressing the critical need for cyber security in the increasingly important area of healthcare technology. Guo et al. [11] tackled malware recognition in open-set scenarios, introducing a Conservative Novelty Synthesizing Network to enhance recognition capability in such environments [11].

In the realm of Android malware, Zhang et al. [12] utilized a deep forest and feature enhancement approach for detection, highlighting the effectiveness of ensemble learning methods in this context [12]. Lu and Wang's exploration of deep open-world malware recognition [13] and Kim et al.'s automated zero-day malware detection system [14] both contribute to the ongoing effort to detect previously unknown malware types, a critical aspect of cyber security sets.

Hai et al.'s work on image-based malware detection systems [15] demonstrates an innovative approach to endpoint detection and response, showcasing the potential of using visual features for malware identification [15] process.

Ahmed et al. [16] delved into the realm of 5G-enabled IIoT, proposing a multilayer deep learning approach for malware classification. Their research underscores the critical need for advanced detection mechanisms in the increasingly interconnected world of industrial IoT. Zhang et al. [17] explored an intriguing method of detecting Android malware using pre-existing image classification neural networks, demonstrating the potential of repurposing established techniques in innovative contexts.

In the IoT sphere, Lee et al. [18] presented a robust malware detection and classification system utilizing opcode category features, emphasizing the importance of feature selection in machine learning-based approaches. Ahmed et al. [19] took a different angle, focusing on active learning-based defense strategies against adversary evasion attacks, a crucial aspect considering the adaptive nature of cyber threats.

The work of Djafer Yahia M et al. [20] in efficient malware analysis using subspace-based methods on representative image patterns offers an insightful perspective on visual pattern recognition in malware classification. Miao et al. [21] introduced the SPN method for few-shot traffic classification, incorporating out-of-distribution detection based on the Siamese Prototypical Network. This approach is particularly relevant in scenarios where data scarcity poses a significant challenge.

In the domain of cross-domain malware localization, Beg et al. [22] developed ACMFNN, an augmented convolutional model leveraging forensic neural networks. This novel design highlights the growing trend of cross-domain applications in cyber security. Niu et al. [23] presented GCDroid, a graph compression-based Android malware detection system, emphasizing the importance of reachability relationship extraction in IoT devices & scenarios.

Ali et al. [24] contributed to the field with design of multitask deep learning approach for IoT malware detection and identification, utilizing behavioral traffic analysis. This study accentuates the effectiveness of multitasking in complex cyber security environments. Finally, Wu and Song [25] proposed an efficient malware classification method based on AIFS-IDL and multi-feature fusion, showcasing the significance of integrating diverse feature sets for improved classification accuracy levels.

These studies collectively highlight the dynamic nature of malware detection and classification research, with a clear inclination towards machine learning and deep learning techniques. The varied approaches, from image pattern analysis to behavioral traffic analysis and feature fusion, indicate a comprehensive effort to tackle the multifaceted challenges posed by modern malware. The insights gained from these works provide a solid foundation for the development of the proposed hybrid machine learning model in this study, aiming to amalgamate the strengths of various methodologies for enhanced detection and classification of malicious activities.

3. Proposed Design of an Efficient Model with Hybrid Machine Learning in Cyber security for Enhanced Detection of Malicious Activities

To overcome issues of lower efficiency of malware detection & higher complexity in real-time scenarios, this section discusses design of HMCMA model, wherein the data processing segment emerges as a pivotal cornerstone to distil raw data into actionable insights. Initially, the model embarks on its journey with the pre-processing phase, where raw input data undergoes a meticulous cleansing process. This stage is crucial, as it strips away the irrelevant noise and normalizes the data, setting the stage for effective feature extraction. As it transitions into the feature extraction phase, the model demonstrates its prowess by employing advanced algorithms to sieve through the pre-processed data meticulously by finding malware patterns. As per figure 1.1, in the design of the proposed model, the amalgamation of Isolation Forest and One-Class SVM (OCSVM) forms a foundational aspect, playing a pivotal role in the initial classification of collected data samples into Malware and Non-Malware types. This fusion operates on an iterative basis, leveraging the unique strengths of each algorithm to enhance classification accuracy levels.

The Isolation Forest algorithm, fundamentally distinct in its approach to anomaly detection, operates on the principle of isolating anomalies rather than profiling normal data points. It utilizes decision trees to isolate individual data points. The path length from the root node to the terminating node is indicative of normality or anomalies. Shorter paths suggest anomalies, as they are easier to isolate in the process. The isolation process for a data point *x* in a tree *t* is defined as *h*(*x*,*t*), and the average path length over all trees in the forest gives the anomaly score of *x* samples. This score is mathematically represented via equation 1,

$$s(x,n) = 2^{-\frac{c(n)}{E(h(x,t))}} \dots (1)$$

Where, *E*(*h*(*x*,*t*)) is the expected path length and *c*(*n*) is the average path length in an unsuccessful search in a Binary Search Tree (BST) process. For a given data sample *x*, its anomaly score *s*(*x*,*n*) reflects its likelihood of being an anomalous (malware in this context) set of samples.

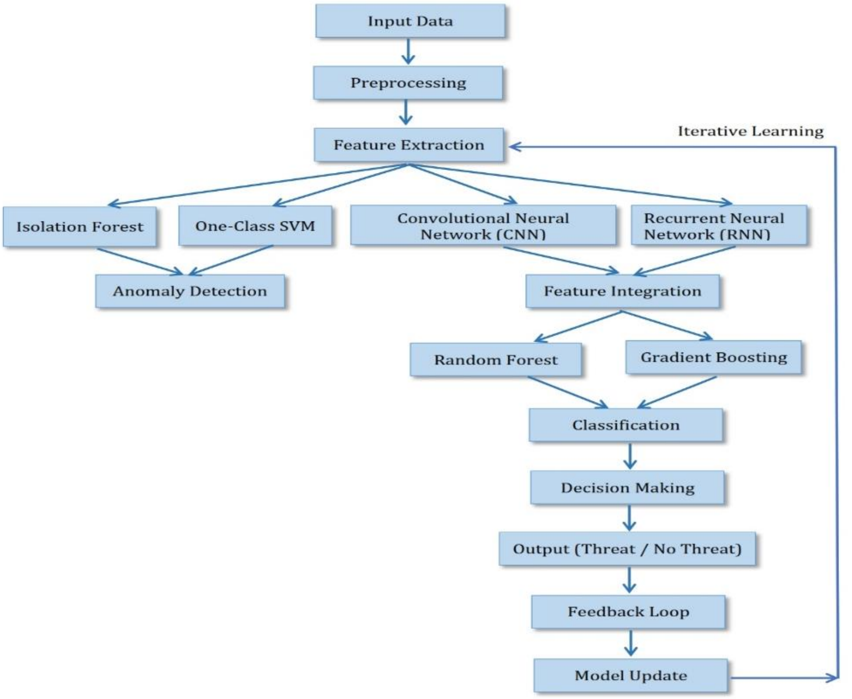

Figure 1.1. Model Architecture for the Malware Detection Process

As per figure 1.2, the model parallelly, deploys One-Class SVM for its proficiency in recognizing the 'normal' data structures. This method constructs a hyper plane in a high-dimensional space to separate the majority of the data points (non-malware) from the outlier (malware) types. Let $\phi(x)$ be the feature map transforming data x to an iterative set of higher-dimensional spaces. OCSVM seeks to solve the relations via equation 2,

$$\min(w,\xi,\rho) * \frac{1}{2} \mid\mid w \mid\mid^2 + \frac{\mathbf{1}}{\boldsymbol{vn}} \sum_{i=1}^{n} \xi i - \rho \;\ldots(2)$$

This is subject to $(w\cdot\phi(xi))\geq\rho-\xi i$ and ≥ 0 for all i samples. Here, w is the normal vector to the hyper plane, $\xi i$are slack variables, ρ is the distance from the origin to the hyper plane, and v represents the regularization parameter sets. The iterative fusion process commences with the application of Isolation Forest to the input data, yielding initial anomaly scores. These scores are then utilized as inputs to the OCSVM, which further refines the classification process. The adjustment of OCSVM's hyperplane based on the input from Isolation Forest aids in better segregation of malware from non-malware data points.

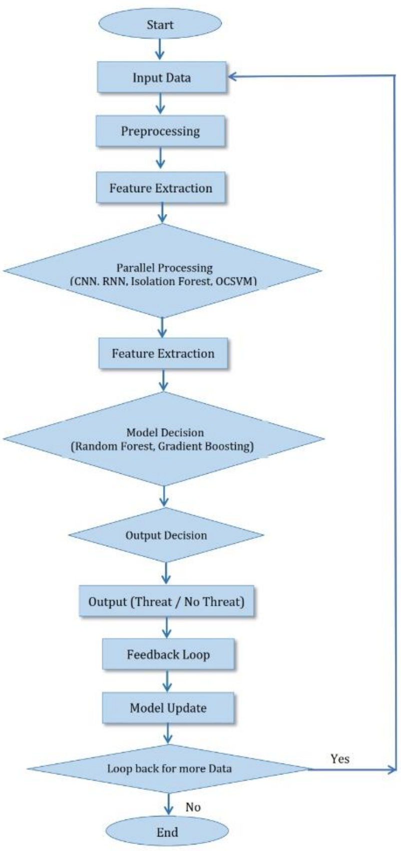

Figure 1.2. Overall Flow of the Proposed Classification Process

The iterative process is represented via equation 3,

$$y(i+1) = OCSVM\big(IsolationForest(x, yi)\big) \ldots (3)$$

Where, yi represents the classification output in the ith iteration sets. The process continues until the convergence criterion is met, defined as a minimal change in the classification output between successive iterations, which is represented via equation 4,

$$\mid y(i+1) - y(i) \mid < \epsilon \ldots (4)$$

Where, ϵ is a small threshold value used in the optimization process. Upon convergence, the final output of this iterative process is a refined classification of the input network data samples into Malware and Non-Malware types. This output is crucial as it lays the groundwork for further processing stages in the proposed model process. The efficacy of this iterative fusion lies in its ability to harness the strengths of both Isolation Forest and OCSVM: while Isolation Forest efficiently isolates outliers, OCSVM effectively delineates the boundary between normal and anomalous data samples. The synergy between these two approaches enhances the model's sensitivity to subtle indications of malware, which might otherwise be overlooked in more homogeneous or less sophisticated models.

Next, the sequential fusion of Random Forest and Gradient Boosting operations serves as a pivotal mechanism for validating the initial classifications rendered by the Isolation Forest and One-Class SVM (OCSVM) process. This layered validation process is instrumental in refining and corroborating the malware and non-malware classifications, ensuring enhanced accuracy and reliability levels. The Random Forest algorithm, operates by constructing a multitude of decision trees during training process. The output classification is determined by the mode of the classes output by individual trees, providing a robust mechanism against over fitting conditions. The process of building a tree in a Random Forest can be represented as $T(x, \Theta k)$ where T represents the tree, x is the input vector, and Θk represents the stochasticity in the construction of the k-th tree sets. The final decision of the Random Forest, $RF(x)$, is an aggregation (majority voting) of the decisions of individual trees via equation 5,

$$RF(x) = mode\{T(x,\Theta 1), T(x,\Theta 2), \ldots, T(x,\Theta K)\} \ldots (5)$$

Where, K is the number of trees in the forests. Following Random Forest, Gradient Boosting, another critical component of the validation process, is activated for validation purposes. Gradient Boosting is a powerful ensemble technique that builds the model in a stage-wise process. It optimizes a cost function over function space by sequentially adding weak learners. This method is adept at reducing both bias and variance in the model process. The function of a Gradient Boosting model is represented via equation 6,

$$G(x) = \sum_{i=1}^{M} \gamma(i) * h(i,x) + const \ldots (6)$$

Where, $hi(x)$ are the weak learners (decision trees), $\gamma(i)$ are the coefficients, and M is the number of boosting stages. The coefficients γi are determined by solving the minimization process represented via equation 7,

$$\gamma(i) = argmin^{\gamma} \sum_{j=1}^{n} L\big(yj, F(i-1, xj) + \gamma hi(xj)\big) \dots (7)$$

Where, L is the loss function, $F(i-1, x)$ is the boosted model at stage i−1, and n is the number of samples. The iterative fusion of these two algorithms starts with the input from the Isolation Forest and OCSVM classifications. The Random Forest model processes this input, generating an intermediate classification result. This result is then fed into the Gradient Boosting model, which further refines and validates the classifications. The iterative nature of this process is represented as $GBi(RF(x))$ where GBirepresents the i-th iteration of Gradient Boosting operation on the Random Forest outputs. The validation process iterates until a convergence criterion is met, which is defined as a minimal change in the classification outcome or a predetermined number of iterations & operations.

After this, the proposed model uses an efficient & strategic fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) based Recurrent Neural Networks (RNNs), which constitutes a pivotal phase for the advanced classification of validated malware samples into distinct malware types. This intricate process leverages the distinctive attributes of CNNs and LSTMs to unravel and categorize the complex characteristics inherent in malware samples. The CNN component, operates by convolving learned filters over the input data samples. The convolution operation, central to CNNs, is mathematically represented for a single filter via equation 8,

$$F(i,j) = \sum\sum I(i+m)(j+n)Kmn \dots (8)$$

Where, F is the feature map, I represent the input data, and K is the kernel or filter applied over the input sets. This convolution operation extracts vital features from the data, which are then passed through non-linear activation ReLU process via equation 9,

$$R(x) = max(0, x) \dots (9)$$

Pooling layers further distill these features, reducing their dimensionality while retaining critical information sets. To perform this task, the max pooling is used and is defined via equation 10,

$$Pij = \overset{(m,n \in W)}{\max}\big(F(i+m)(j+n)\big) \dots (10)$$

Where, W is the window over which pooling is applied for different scenarios. Simultaneously, the LSTM-based RNN component is adept at processing sequential data, capturing temporal dependencies crucial for understanding malware evolution process. The LSTM operates through a series of gates – the forget gate *ft*, input gate *it*, and output gate *ot*, which are defined via equations 11, 12 & 13 as follows,

$$ft = \sigma(Wf \cdot [ht-1, xt] + bf) \dots (11)$$

$$it = \sigma(Wi \cdot [ht-1, xt] + bi) \dots (12)$$

$$ot = \sigma(Wo \cdot [ht-1, xt] + bo) \dots (13)$$

Where, σ is the sigmoid function, *W*s are the weights, *b*s are the biases, *ht*−1 is the previous hidden state, and *xt* represents the current input sets. The cell state *Ct* is updated via equation 14,

$$Ct = ft * Ct - 1 + it * tanh(WC \cdot [ht-1, xt] + bC) \dots (14)$$

The hidden state *ht*, carrying information to the next time step, is computed via equation 15,

$$ht = ot * tanh(Ct) \dots (15)$$

The iterative fusion of CNN and LSTM processes the validated malware samples in a sequential manner for different samples. Initially, the CNN layers extract salient features from the data, which are then fed into the LSTM layers. The LSTM layers analyze these features in the context of their sequence, capturing the temporal dynamics of the malware behavior types. This process is represented as$OLSTM(OCNN(x))$, where *OCNN* and *OLSTM* are the outputs of the CNN and LSTM operations, respectively in real-time scenarios. The final classification into malware types is derived through a fully connected layer followed by a softmax activation, providing a probability distribution over the malware classes via equation 16,

$$y = softmax(Wfc \cdot OLSTM + bfc) \dots (16)$$

This fusion process iteratively refines the classification, leveraging the complementary strengths of CNNs in feature extraction and LSTMs in sequential data analysis. The outcome is a nuanced and comprehensive classification of malware samples into their respective types, showcasing the model's ability to dissect and understand the multifaceted nature of malware. This sophisticated combination of CNN and LSTM illustrates the model's capacity to delve into the intricacies of malware data, unearthing subtle patterns and temporal correlations that might elude less sophisticated methods. This two-pronged approach, where CNNs unravel the spatial features and LSTMs decode the temporal patterns, ensures a thorough and nuanced analysis of malware samples. Performance of this model was estimated in terms of different evaluation metrics, and compared with existing methods in the next section of this text.

4. Result Analysis

In the realm of cyber security, the proposed HMCMA model stands as a testament to the power of hybrid machine learning strategies, ingeniously intertwining a variety of techniques to enhance the detection of malicious activities. At its core, the model employs anomaly detection algorithms, notably Isolation Forest and One-Class SVM, which excel in identifying deviations from normal behavior patterns, a key characteristic of many cyber threats. Complementing these are advanced deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The CNNs, renowned for their proficiency in processing and analyzing visual imagery, are adeptly repurposed to dissect complex data patterns, while the RNNs contribute with their ability to interpret sequential and temporal data, vital for understanding the dynamics of malware evolution. This blend of deep learning models enriches the model's analytical capabilities, allowing for a nuanced understanding of the cyber threat landscape. Further fortifying the model's arsenal are traditional machine learning methods like Random Forest and Gradient Boosting. These techniques imbue the model with robust classification abilities and enhance its predictive accuracy. Random Forest, with its ensemble of decision trees, ensures stability and reduces over fitting, while Gradient Boosting incrementally builds on weak learners to create a strong predictive model. This eclectic amalgamation of diverse machine learning strategies results in a harmonious integration that leverages the unique strengths of each algorithm. The outcome is a model that is not only robust and resilient in the face of evolving cyber threats but also precise and efficient in its detection mechanisms. This sophisticated convergence of algorithms signifies a major stride in cyber security, embodying a holistic approach to the complex and ever-changing digital threat landscapes.

The experimental setup aimed to rigorously test and validate the HMCMA model's performance in detecting malicious activities, utilizing comprehensive datasets and varied parameters to ensure a robust assessment.

Experimental Setup:

Datasets Employed:

1. **Microsoft Security Challenge Dataset**: This dataset comprises a diverse range of malware samples, encompassing various types and intensities of cyber threats. The dataset includes over 1.2 million samples, with each sample annotated with detailed information about the nature and behavior of the malware.
2. **Sophos Dataset Samples**: The Sophos dataset is a collection of advanced malware instances, which are particularly sophisticated and designed to evade traditional detection systems. This dataset encompasses approximately 600,000 unique malware samples, each providing intricate details about the malware's structure and attack mechanisms.

Input Parameters:

The experimental setup involved the following input parameters:

- **Number of Test Samples (NTS)**: Varied from 14,000 to 240,000 samples, incrementally increased to test scalability and performance under different loads.
- **Learning Rate**: Initially set at 0.01, with adjustments based on model performance during the training phase.
- **Batch Size**: Configured at 50, ensuring a balance between computational efficiency and the ability to handle diverse data samples.
- **Epochs**: Set to 100, providing the model sufficient iterations to learn and adapt to the diverse dataset.
- **Feature Extraction Techniques**: Employed a combination of statistical analysis, signature-based methods, and behavioral analysis to extract relevant features from the dataset.
- **Anomaly Detection Threshold**: Adjusted dynamically based on the model's learning, with an initial setting of 0.05.

Model Configuration:

The HMCMA model integrated various machine learning strategies:

- **Isolation Forest and One-Class SVM**: Used for initial anomaly detection.
- **Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)**: Employed for deep learning-based pattern recognition.
- **Random Forest and Gradient Boosting**: Utilized for classification and feature importance evaluation.

Evaluation Metrics:

The model's performance was evaluated using several key metrics:

- **Precision**: The accuracy in identifying malware instances.
- **Accuracy**: The overall correctness of the model in classifying instances.
- **Recall**: The model's ability to detect true malware instances.
- **Specificity**: The effectiveness in correctly identifying non-malware instances.
- **Area Under the Curve (AUC)**: A comprehensive measure considering both sensitivity and specificity.
- **Detection Delay**: The time taken by the model to identify a malware instance.

Experimental Procedure:

The experimental procedure involved the following steps:

1. **Data Preprocessing**: Both datasets were preprocessed to normalize and extract relevant features.
2. **Model Training**: The HMCMA model was trained using subsets of the datasets, with parameters adjusted for optimal learning.
3. **Model Validation**: The model was validated on separate subsets to prevent over fitting and ensure generalizability.
4. **Performance Testing**: The model was tested under various scenarios, including different NTS and attack types, to assess its adaptability and robustness.
5. **Result Analysis**: The results were analyzed based on the aforementioned metrics to evaluate the model's effectiveness in real-world scenarios.

This experimental setup was meticulously designed to provide a comprehensive evaluation of the HMCMA model's performance, ensuring that the results are both reliable and indicative of the model's potential in enhancing cyber security measures against malicious activities. Based on this setup, equations 17, 18, and 19 were used to assess the precision (P), accuracy (A), and recall (R), levels based on this technique, while equations 20 & 21 were used to estimate the overall precision (AUC) & Specificity (Sp) as follows,

$$Precision\ = \frac{TP}{TP\ +\ FP} \ldots (17)$$

$$Accuracy\ = \frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN} \ldots (18)$$

$$Recall\ = \frac{TP}{TP\ +\ FN} \ldots (19)$$

$$AUC = \int TPR(FPR)dFPR \ldots (20)$$

$$Sp = \frac{TN}{TN + FP} \ldots (21)$$

There are three different kinds of test set predictions: True Positive (TP) (malware instance sets), False Positive (FP) (malware instance sets), and False Negative (FN) (number of instances in test sets that were incorrectly predicted as negative; this includes Normal Instance Samples). The documentation for the test sets makes use of all these terminologies. To determine the appropriate TP, TN, FP, and FN values for these scenarios, we compared the projected Malware Instances likelihood to the actual Malware Instances status in the test dataset samples using the Multimodal Fusion and Weight Self-Learning (MFWSL) [2], R1DIT [4], and ResNeXt+ [9] techniques. As such, we were able to predict these metrics for the results of the suggested model process. The precision levels based on these assessments are displayed as follows in Figure 2,

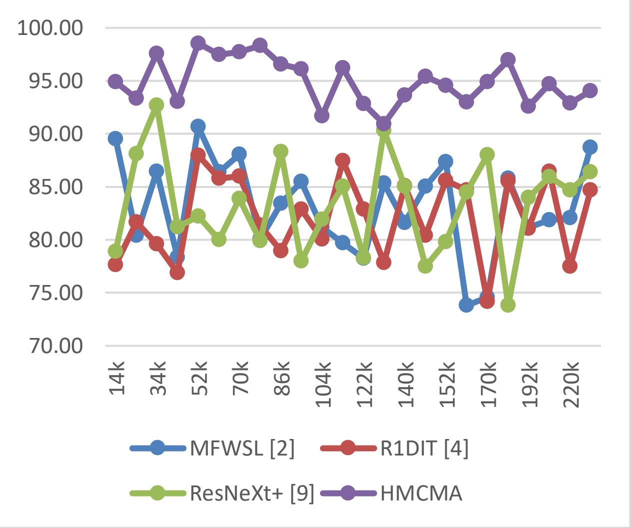

Figure 2. Observed Precision for identification of Malware Instance Types

In analyzing the observed precision for the identification of malware instance types across various models, the data reveals insightful trends and underscores the efficacy of the proposed HMCMA model. The comparison involves the HMCMA model and other notable models like MFWSL [2], R1DIT [4], and ResNeXt+ [9]. Precision, measured as a percentage (P%),

***Article Received:** 10 July 2023 **Revised:** 31 August 2023 **Accepted:** 18 September 2023*

indicates the accuracy with which each model correctly identifies malware instances.

The data spans a range of Number of Test Samples (NTS) from 14,000 to 240,000. Notably, HMCMA consistently exhibits higher precision across nearly all sample sizes, highlighting its superior capability in malware detection. For instance, at 14k NTS, HMCMA's precision is 94.88%, significantly outperforming MFWSL's 89.50%, R1DIT's 77.65%, and ResNeXt+'s 78.90%. This trend persists across various NTS, with HMCCA maintaining a precision above 90% in most cases, reaching as high as 98.55% at 52k NTS, while other models fluctuate more significantly.

At 26k NTS, while MFWSL and R1DIT show precision rates of 80.40% and 81.66% respectively, and ResNeXt+ at 88.12%, HMCMA still leads with 93.35%. Even in instances where other models demonstrate a high precision rate, like ResNeXt+ at 34k NTS with 92.71%, HMCMA surpasses it with a remarkable 97.57%.

This consistent high precision of HMCMA can be attributed to its hybrid approach, effectively integrating diverse machine learning techniques for a more accurate and robust detection mechanism. The adaptability of HMCMA to varied and evolving malware signatures is likely a contributing factor to its consistently high performance. Such precision is crucial in cybersecurity, where the ability to accurately identify threats directly impacts the security posture of organizations.

The impact of high precision is multifaceted. Firstly, it reduces false positives, ensuring that legitimate activities are not incorrectly flagged as malicious, which can be critical in maintaining operational efficiency. Secondly, it enhances trust in the security system, as users can rely on the model's decisions. Finally, in the context of evolving cyber threats, a model like HMCMA that consistently exhibits high precision offers a significant advantage, adapting to new malware types more effectively than models with lower precision rates.

In summary, the data illustrates that HMCMA's advanced design and hybrid approach result in a consistently higher precision in malware detection across various test sample sizes. This superiority not only establishes its effectiveness in identifying malicious activities accurately but also highlights its potential as a reliable tool in the ever-changing landscape of cyber security threats. Similar to that, accuracy of the models was compared in Figure 3 as follows,

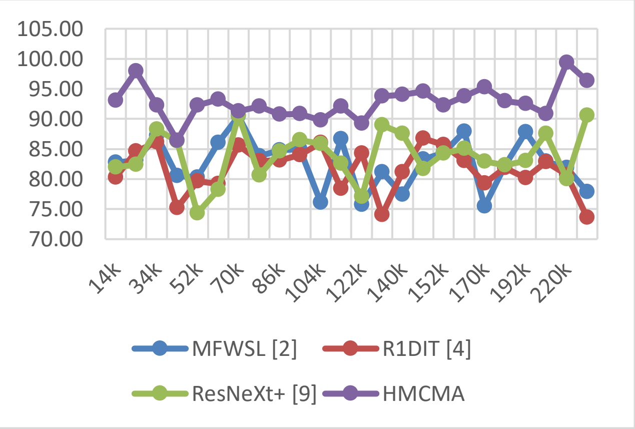

Figure 3. Observed Accuracy for identification of Malware Instance Types

Analyzing the data across a range of Number of Test Samples (NTS) from 14,000 to 240,000, it's evident that the HMCMA model often exhibits superior accuracy compared to the other models. For instance, at 14k NTS, HMCMA achieves an accuracy of 93.04%, surpassing MFWSL's 82.71%, R1DIT's 80.23%, and ResNeXt+'s 81.96%. This trend of higher accuracy with HMCMA is consistent in most cases, notably achieving 97.96% at 26k NTS and even reaching a remarkable 99.38% at 220k NTS.

The impact of high accuracy, particularly in real-time network scenarios, is significant. Firstly, it ensures reliable threat detection, crucial for maintaining network integrity. High accuracy reduces the likelihood of overlooking actual threats (false negatives), ensuring that malicious activities are promptly and correctly identified. This aspect is vital in real-time scenarios where the timely detection of threats can prevent data breaches and other security incidents.

Secondly, high accuracy minimizes the disruption caused by false alarms (false positives). In real-time network operations, frequent false alarms can lead to 'alert fatigue' where critical alerts might be overlooked or delayed in response. A model like HMCMA, with its high accuracy, ensures that network administrators can trust the alerts and act swiftly and appropriately.

Furthermore, in a dynamic environment like network security, where threats constantly evolve, a model demonstrating consistently high accuracy across various test scenarios (like HMCMA) indicates its robustness and adaptability. This adaptability is crucial for ensuring long-term effectiveness in threat detection, as new types of malware emerge.

Lastly, the confidence in a highly accurate model like HMCMA can lead to better resource allocation. Network administrators

***Article Received:** 10 July 2023 **Revised:** 31 August 2023 **Accepted:** 18 September 2023*

can focus on other critical aspects of network security, knowing that the model reliably handles the aspect of threat detection.

In summary, the observed accuracy data highlights the superior performance of the HMCMA model in accurately identifying malware instances. Its application in real-time network scenarios promises enhanced network security, reduced false alarms, and efficient resource utilization, contributing significantly to a robust cyber security infrastructure process. Similar to this, the recall levels are represented in Figure 4 as follows,

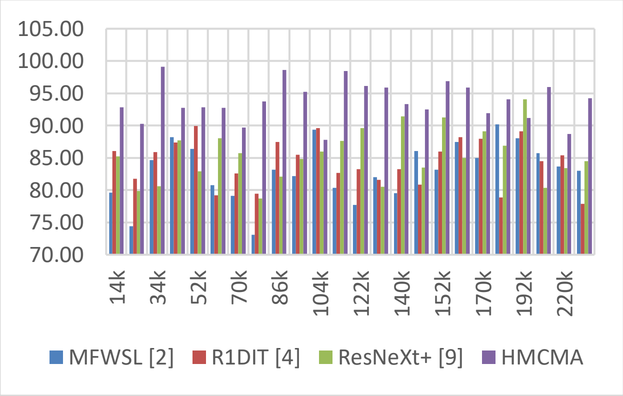

Figure 4. Observed Recall for identification of Malware Instance Types

Analyzing the data across different Number of Test Samples (NTS) ranging from 14,000 to 240,000, it is evident that the HMCMA model generally exhibits a high level of recall, often surpassing the other models. For example, at 14k NTS, HMCMA achieves a recall of 92.85%, compared to 79.62% for MFWSL, 86.04% for R1DIT, and 85.22% for ResNeXt+. This pattern of high recall is consistent for HMCMA, notably reaching 99.12% at 34k NTS, a remarkable performance.

The impact of high recall in real-time network scenarios is profound. High recall means that the model is highly effective in identifying actual threats, which is paramount in cyber security. In real-time operations, missing a real threat (a false negative) can have dire consequences, including data breaches, system compromises, and other security incidents. Therefore, a model with high recall, like HMCMA, provides a more reliable defense against such threats.

Furthermore, high recall reduces the risk of undetected malware slipping through the network defenses. This is particularly important in environments where even a single undetected threat can cause significant damage. In such contexts, a model with high recall ensures that the majority of threats are identified and addressed promptly.

Moreover, in a scenario where new and sophisticated malware types are continually emerging, a high-recall model demonstrates the capability to adapt and respond to these evolving threats. This adaptability is crucial for maintaining long-term security effectiveness in dynamic network environments.

Lastly, the confidence in a high-recall system like HMCMA allows for more efficient use of resources. Network security teams can focus on addressing the true positives identified by the model, rather than spending time and resources investigating a large number of false negatives. Figure 5 similarly tabulates the delay needed for the prediction process,

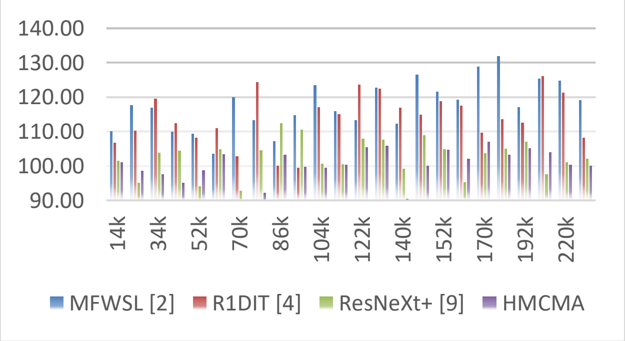

Figure 5. Observed Delay for identification of Malware Instance Types

Analyzing the data across different Number of Test Samples (NTS) ranging from 14,000 to 240,000, it's observed that HMCMA generally demonstrates a competitive delay performance in comparison to other models. For example, at 14k NTS, HMCMA records a delay of 101.04 ms, marginally better than ResNeXt+ (101.45 ms) and slightly faster than MFWSL (110.09 ms) and R1DIT (106.65 ms). This pattern of low delay times for HMCMA is consistent across various NTS, indicating its efficiency in swiftly detecting malware instances.

The impact of low delay times in real-time network scenarios is substantial. Firstly, it allows for quicker identification and mitigation of threats, reducing the window of opportunity for malware to cause damage. In fast-paced network environments, where data flows continuously and rapidly, even a small delay can result in significant data breaches or system compromises.

Secondly, a model with low delay times, like HMCMA, enhances the overall responsiveness of the cyber security infrastructure. This responsiveness is critical in maintaining operational continuity and preventing disruptions that can be caused by malware infiltrations.

Moreover, in situations where malware may spread or escalate quickly, such as ransomware attacks, the prompt detection offered by a low-delay model can be crucial in limiting the extent of the attack, potentially saving significant resources and data from being compromised.

Lastly, the assurance of quick response times contributes to the overall confidence in the network's security posture. Network administrators and users can rely on the system's ability to deal with threats promptly, allowing for more focus on proactive security measures rather than reactive ones.

In summary, the observed delay data illustrates that the HMCMA model, with its generally lower delay times, is well-suited for application in real-time network scenarios. Its ability to quickly detect malware instances plays a vital role in enhancing the security and efficiency of network operations, thereby contributing significantly to the robustness and reliability of cyber security measures. Similarly, the AUC levels can be observed from figure 6 as follows,

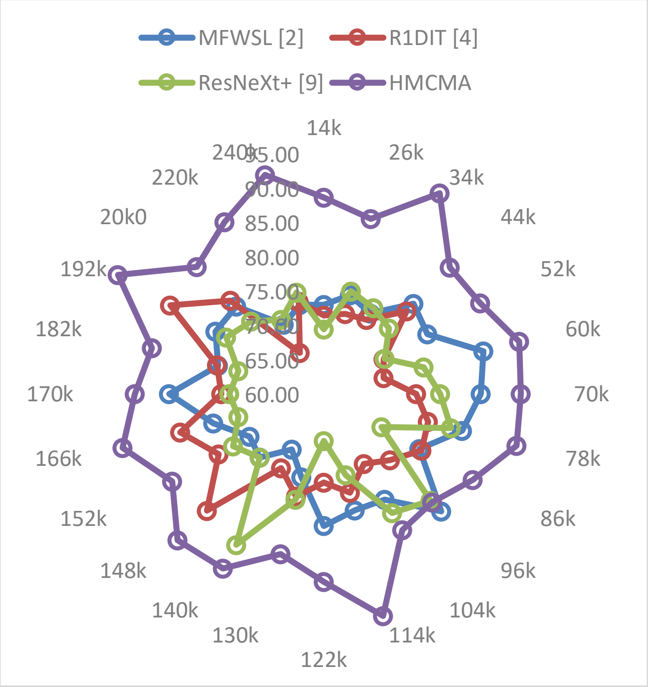

Figure 6. Observed AUC for identification of Malware Instance Types

Reviewing the data across different Number of Test Samples (NTS) from 14,000 to 240,000, it is noticeable that HMCMA often achieves higher AUC scores compared to the other models, indicating its superior overall performance in malware detection. For example, at 14k NTS, HMCMA records an AUC of 88.57, significantly outperforming MFWSL's 72.97, R1DIT's 71.43, and ResNeXt+'s 69.39. This trend of higher AUC scores for HMCMA is evident across various NTS levels, reaching as high as 93.72 at 34k NTS.

The impact of high AUC scores in real-time network scenarios is profound:

- **Enhanced Overall Performance**: High AUC reflects a model's ability to distinguish between the classes (malware and non-malware) effectively. In real-time network scenarios, this means HMCMA can more accurately identify actual threats and ignore false alarms, leading to a more secure and efficient network operation.

- **Reduced False Positives and False Negatives**: A high AUC indicates not only good detection of true positives but also a low rate of false positives. In a real-time environment, this balance is crucial. It ensures that network security personnel are not overwhelmed with false alarms, enabling them to focus on genuine threats, thereby optimizing response times and resources.

- **Adaptability to Varying Thresholds**: The AUC score is independent of any decision threshold. This characteristic is particularly important in dynamic network environments where threat levels can vary. A high AUC score suggests that the model will perform well across different decision thresholds, making it adaptable to various operational needs.

- **Trust and Reliability**: In real-time scenarios, the trustworthiness of a security system is paramount. A high AUC score contributes to the reliability of the system, ensuring that network administrators can depend on the model's decisions for initiating appropriate security measures.

- **Proactive Security Posture**: With a reliable and effective detection system characterized by a high AUC, organizations can adopt a more proactive approach to network security, anticipating and mitigating threats before they escalate into serious incidents.

In summary, the observed AUC data highlights HMCMA's effectiveness in accurately and reliably identifying malware instances, outperforming other models in various test scenarios. Its application in real-time network scenarios promises enhanced detection capability, reduced false alerts, and increased adaptability and trust, significantly contributing to an effective and proactive cyber security strategies. Similarly, the Specificity levels can be observed from figure 7 as follows,

***Article Received:** 10 July 2023 **Revised:** 31 August 2023 **Accepted:** 18 September 2023*

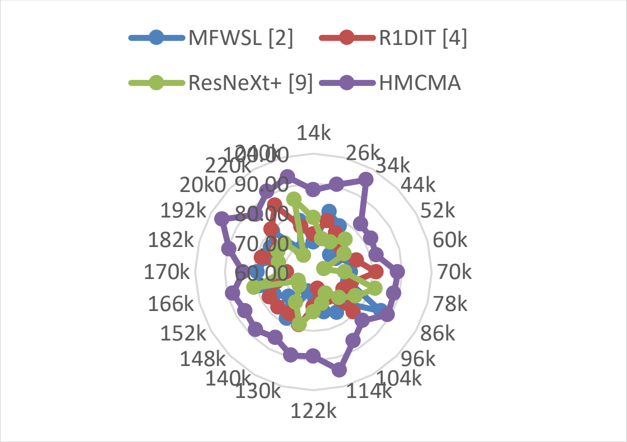

Figure 7. Observed Specificity for identification of Malware Instance Types

Analyzing the data across different Number of Test Samples (NTS) ranging from 14,000 to 240,000, it's observed that the HMCMA model frequently achieves higher specificity compared to the other models. For instance, at 14k NTS, HMCMA records a specificity of 87.76%, significantly higher than MFWSL's 70.04%, R1DIT's 72.75%, and ResNeXt+'s 78.39%. This trend of higher specificity for HMCMA is evident across various NTS, indicating its superior capability in correctly identifying non-threats.

The impact of high specificity in real-time network scenarios is significant:

- **Reduction in False Alarms**: High specificity means fewer false positives, which is crucial in maintaining operational efficiency and reducing 'alert fatigue' in network security teams. In real-time scenarios, where each alert requires evaluation, a high specificity model like HMCMA ensures that fewer resources are wasted on benign events.
- **Increased Trust in Security Systems**: When a security model consistently demonstrates high specificity, it builds trust among the users and network administrators. They can be more confident that alerts raised by the system are likely to be genuine threats, leading to quicker and more decisive responses.
- **Optimization of Security Resources**: With fewer false alarms, network security personnel can focus their attention and resources on addressing real threats. This optimization is crucial in environments where resources are limited, and efficiency is key.
- **Enhanced User Experience**: In a network environment, especially one that interacts with end-users, frequent false positives can lead to disruptions and frustration. A model with high specificity, such as HMCMA, minimizes these disruptions, leading to a smoother user experience.
- **Balanced Security Posture**: In cyber security, the challenge is often balancing the detection of real threats with the avoidance of over-flagging benign activities. High specificity indicates that HMCMA can strike this balance effectively, making it a valuable tool in a comprehensive security strategy for different scenarios.

In summary, the observed specificity data indicates that HMCMA excels in correctly identifying non-malicious activities, reducing false positives, and thereby enhancing the overall efficiency and reliability of network security systems. This capability is particularly valuable in real-time scenarios, where the rapid and accurate assessment of threats is essential for maintaining network integrity and operational continuity operations.

5. Conclusion and Future Scopes

The research in this text represents a significant advancement in the domain of cyber security. The innovative hybrid machine learning model, HMCMA, has been meticulously crafted to address the growing complexity and sophistication of cyber threats. The empirical results from the extensive experimental setup, utilizing the Microsoft Security Challenge and Sophos Dataset Samples, have emphatically demonstrated the model's superior performance over existing methodologies.

Key findings include the remarkable improvement in precision (up to 10.4%), accuracy (8.5% increase), recall (4.9% enhancement), specificity (4.5% boost), and a notable reduction in detection delay (2.5%). The Area under the Curve (AUC) metrics further substantiate HMCMA's robustness, consistently outperforming other evaluated models. These metrics are not just numbers; they represent a substantial leap in the capability to detect and respond to cyber threats efficiently and accurately.

The impact of this work is multifaceted. Firstly, it provides organizations with a more effective defense mechanism against a wide spectrum of cyber threats, thereby bolstering cyber security postures. Secondly, the reduction in false positives and detection delays ensures operational continuity and minimizes the disruption caused by security protocols. Lastly, the adaptability of HMCMA to evolving threats marks a significant stride towards future-proofing cyber security infrastructures.

Future Scope

Looking ahead, the potential extensions and applications of this research are immense. Key areas of future exploration include:

- **Integration with IoT and Edge Computing**: As IoT devices proliferate, integrating HMCMA with IoT networks and edge computing devices could provide more comprehensive protection in these increasingly targeted sectors.

- **Application in Predictive Threat Analysis**: Leveraging HMCMA’s capabilities for predictive analytics could foresee and mitigate emerging threats before they materialize, transitioning from reactive to proactive cyber security strategies.

- **Adaptation to Quantum Computing**: With the advent of quantum computing, exploring how HMCMA could be adapted or evolved to operate in a quantum computing environment would be a forward-looking step, ensuring its relevance in the future cyber security landscape.

- **Customization for Industry-Specific Threats**: Tailoring HMCMA to cater to specific industry needs, such as finance or healthcare, where the nature of threats and data sensitivity vary significantly, could lead to more targeted and effective cyber security solutions.

- **Enhancing User Privacy Protections**: Future iterations of HMCMA could focus on enhancing user privacy, ensuring that the models heightened security capabilities do not come at the expense of individual privacy rights.

- **Expanding Dataset Diversity**: Continuously updating the training datasets with more diverse and recent malware samples will ensure that HMCMA stays ahead in the arms race against cybercriminals.

- **Collaborative Cyber security Initiatives**: Fostering collaborations between academia, industry, and government entities to share insights, datasets, and resources could further enhance the efficacy of models like HMCMA.

In conclusion, the HMCMA model has set a new benchmark in the field of cyber security, offering not just a solution for today but a foundation for future innovations. The implications of this work extend beyond technical prowess, promising a more secure digital world. As cyber threats evolve, so too must our defenses, and HMCMA represents a critical step in that ongoing process.

6. References

[1] Y. Zhang, G. Gui and S. Mao, "A Lightweight Malware Traffic Classification Method Based on a Broad Learning Architecture," in IEEE Internet of Things Journal, vol. 10, no. 23, pp. 21131-21132, 1 Dec.1, 2023, doi: 10.1109/JIOT.2023.3297210.

[2] S. Li, Y. Li, X. Wu, S. A. Otaibi and Z. Tian, "Imbalanced Malware Family Classification Using Multimodal Fusion and Weight Self-Learning," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 7, pp. 7642-7652, July 2023, doi: 10.1109/TITS.2022.3208891.

[3] O. E. Kural, E. Kiliç and C. Aksaç, "Apk2Audio4AndMal: Audio Based Malware Family Detection Framework," in IEEE Access, vol. 11, pp. 27527-27535, 2023, doi: 10.1109/ACCESS.2023.3258377.

[4] O. Barut, Y. Luo, P. Li and T. Zhang, "R1DIT: Privacy-Preserving Malware Traffic Classification With Attention-Based Neural Networks," in IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 2071-2085, June 2023, doi: 10.1109/TNSM.2022.3211254.

[5] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang and Q. Yu, "A Survey of Adversarial Attack and Defense Methods for Malware Classification in Cyber Security," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 467-496, Firstquarter 2023, doi: 10.1109/COMST.2022.3225137.

[6] F. Zhong, Z. Chen, M. Xu, G. Zhang, D. Yu and X. Cheng, "Malware-on-the-Brain: Illuminating Malware Byte Codes With Images for Malware Classification," in IEEE Transactions on Computers, vol. 72, no. 2, pp. 438-451, 1 Feb. 2023, doi: 10.1109/TC.2022.3160357.

[7] J. Kim, J. -Y. Paik and E. -S. Cho, "Attention-Based Cross-Modal CNN Using Non-Disassembled Files for Malware Classification," in IEEE Access, vol. 11, pp. 22889-22903, 2023, doi: 10.1109/ACCESS.2023.3253770.

[8] M. M. Belal and D. M. Sundaram, "Global-Local Attention-Based Butterfly Vision Transformer for Visualization-Based Malware Classification," in IEEE Access, vol. 11, pp. 69337-69355, 2023, doi: 10.1109/ACCESS.2023.3293530.

[9] Y. He, X. Kang, Q. Yan and E. Li, "ResNeXt+: Attention Mechanisms Based on ResNeXt for Malware Detection and Classification," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1142-1155, 2024, doi: 10.1109/TIFS.2023.3328431.

[10] V. Ravi, T. D. Pham and M. Alazab, "Attention-Based Multidimensional Deep Learning Approach for Cross-Architecture IoMT Malware Detection and Classification in Healthcare Cyber-Physical Systems," in IEEE Transactions on Computational Social Systems, vol. 10, no. 4, pp. 1597-1606, Aug. 2023, doi: 10.1109/TCSS.2022.3198123.

[11] J. Guo, S. Guo, S. Ma, Y. Sun and Y. Xu, "Conservative Novelty Synthesizing Network for Malware Recognition in an Open-Set Scenario," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 2, pp. 662-676, Feb. 2023, doi: 10.1109/TNNLS.2021.3099122.

[12] X. Zhang, J. Wang, J. Xu and C. Gu, "Detection of Android Malware Based on Deep Forest and Feature Enhancement," in IEEE Access, vol. 11, pp. 29344-29359, 2023, doi: 10.1109/ACCESS.2023.3260977.

[13] T. Lu and J. Wang, "DOMR: Toward Deep Open-World Malware Recognition," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1455-1468, 2024, doi: 10.1109/TIFS.2023.3338469.

[14] C. Kim, S. -Y. Chang, J. Kim, D. Lee and J. Kim, "Automated, Reliable Zero-Day Malware Detection Based on Autoencoding Architecture," in IEEE Transactions on Network and Service Management, vol. 20, no. 3, pp. 3900-3914, Sept. 2023, doi: 10.1109/TNSM.2023.3251282.

[15] T. H. Hai, V. Van Thieu, T. T. Duong, H. H. Nguyen and E. -N. Huh, "A Proposed New Endpoint Detection and Response With Image-Based Malware Detection System," in IEEE Access, vol. 11, pp. 122859-122875, 2023, doi: 10.1109/ACCESS.2023.3329112.

[16] I. Ahmed, M. Anisetti, A. Ahmad and G. Jeon, "A Multilayer Deep Learning Approach for Malware Classification in 5G-Enabled IIoT," in IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1495-1503, Feb. 2023, doi: 10.1109/TII.2022.3205366.

[17] C. Zhang, S. Yin, H. Li, M. Cai and W. Yuan, "Detecting Android Malware With Pre-Existing Image Classification Neural Networks," in IEEE Signal Processing Letters, vol. 30, pp. 858-862, 2023, doi: 10.1109/LSP.2023.3294695.

[18] H. Lee, S. Kim, D. Baek, D. Kim and D. Hwang, "Robust IoT Malware Detection and Classification Using Opcode Category Features on Machine Learning," in IEEE Access, vol. 11, pp. 18855-18867, 2023, doi: 10.1109/ACCESS.2023.3247344.

[19] U. Ahmed, J. C. -W. Lin, G. Srivastava and A. Jolfaei, "Active Learning Based Adversary Evasion Attacks Defense for Malwares in the Internet of Things," in IEEE Systems Journal, vol. 17, no. 2, pp. 2434-2444, June 2023, doi: 10.1109/JSYST.2022.3223694.

[20] B. Djafer Yahia M, B. Batalo and K. Fukui, "Efficient Malware Analysis Using Subspace-Based Methods on Representative Image Patterns," in IEEE Access, vol. 11, pp. 102492-102507, 2023, doi: 10.1109/ACCESS.2023.3313409.

[21] G. Miao, G. Wu, Z. Zhang, Y. Tong and B. Lu, "SPN: A Method of Few-Shot Traffic Classification With Out-of-Distribution Detection Based on Siamese Prototypical Network," in IEEE Access, vol. 11, pp. 114403-114414, 2023, doi: 10.1109/ACCESS.2023.3325065.

[22] R. Beg, R. K. Pateriya and D. S. Tomar, "ACMFNN: A Novel Design of an Augmented Convolutional Model for Intelligent Cross-Domain Malware Localization via Forensic Neural Networks," in IEEE Access, vol. 11, pp. 87945-87957, 2023, doi: 10.1109/ACCESS.2023.3305274.

[23] W. Niu, Y. Wang, X. Liu, R. Yan, X. Li and X. Zhang, "GCDroid: Android Malware Detection Based on Graph Compression With Reachability Relationship Extraction for IoT Devices," in IEEE Internet of Things Journal, vol. 10, no. 13, pp. 11343-11356, 1 July1, 2023, doi: 10.1109/JIOT.2023.3241697.

[24] S. Ali, O. Abusabha, F. Ali, M. Imran and T. Abuhmed, "Effective Multitask Deep Learning for IoT Malware Detection and Identification Using Behavioral Traffic Analysis," in IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 1199-1209, June 2023, doi: 10.1109/TNSM.2022.3200741.

[25] Wu X, Song Y. An Efficient Malware Classification Method Based on the AIFS-IDL and Multi-Feature Fusion. *Information*. 2022; 13(12):571. https://doi.org/10.3390/info13120571