Explainable AI-Based Malware Detection for Protecting Assistive Learning Systems Used by Students with Special Needs

Abdullah Farhan J Alshammari[1]

[1]Assistant Professor, Department of CSE, Yanbu Industrial College, Saudi Arabia.

Email: shammariafj@rcjy.edu.sa, Orcid id# 0009-0006-5215-4724

HOW TO CITE:
Abdullah Farhan J Alshammari (2026). Explainable AI-Based Malware Detection for Protecting Assistive Learning Systems Used by Students with Special Needs. International Journal of Special Education, 41(8s), 843-858

ABSTRACT:
The use of assistive learning technologies, adaptive educational platforms, and digital communication systems has greatly facilitated education in terms of accessibility for students with special needs. The growing use of these technologies, however, has led to a greater risk of being targeted by malware and other cyber threats that can impact the safety of sensitive education information. Sophisticated and evolving malware are often not detected by traditional detection methods, and many deep learning methods are not interpretable. In order to cope with these challenges, the present study presents an Explainable Artificial Intelligence (XAI) based malware detection approach for the security of assistive learning systems. The proposed architecture is based on both structured malware features and deep feature representations derived from convolutional neural networks and learning modules based on Transformer models. A feature refinement mechanism based on attention and a conditional feature fusion strategy are used to improve the discriminative learning and the classification of malware. For better transparency, Shapley Additive Explanations (SHAP) are included to offer interpretable insights into classification decisions. Experimental testing shows that the proposed approach is effective on both the SOMLAP and the EMBER benchmark malware datasets with accuracy (99.65%), precision (1.00), recall (0.996) and F1-score (0.998) better than any conventional machine learning approach. The proposed framework offers a safe, solid, and transparent cybersecurity solution to protect and secure assistive educational systems.

Keywords: Explainable Artificial Intelligence, Special Education, Cybersecurity, Malware Detection, Assistive Learning Systems.

1. Introduction

The use of digital technologies has had a profound impact on the way learning is done today, and especially in special education settings. ALS are tools, such as screen readers, speech generating devices, adaptive communication systems, intelligent tutoring systems, and personalized e-learning systems, that have helped make learning more accessible to students with special needs, thereby promoting independent

learning, communication and participation in schooling. Artificial intelligence, cloud computing, the Internet of Things (IoT), and learning platforms have also enhanced the potential of AT, allowing students to access learning opportunities despite physical, sensory, or cognitive barriers (Algarni & Thayananthan, 2025).

While these progressions have led to positive developments, there have also been many challenges to cybersecurity with the reliance on digital learning infrastructures. More and more educational institutions are keeping sensitive data like student records, individualized education plans (IEPs), therapy reports, behavioral assessments, and communication data. As a result, educational systems are now an attractive target for cybercriminals to gain access to confidential information without permission or to disrupt educational services. Recent research has underscored the escalating risk of smart education ecosystems from malicious attacks, such as malware, ransomware, and data breaches, highlighting the growing criticality of cybersecurity in learning facilities (Ali et al., 2025).

Cyberattacks can have a specific impact on students with special educational needs, as they often heavily depend on assistive technology to communicate, learn and access educational materials. These systems can be compromised and become unavailable and insecure from Malware, which can negatively impact learning activities and outcomes. According to Awang et al. (2024), special needs students' cybersecurity awareness is still at a moderate level, highlighting the importance of increasing the level of technological protection and awareness campaigns. Likewise, Guduru et al. (2026) emphasized the need to secure individual educational and therapy records, citing the potential impacts of security breaches on the continuity of education and therapy for learners with disabilities.

Malware is one of the most prevalent and dynamic types of cyber threats. Conventional methods of malware detection focus mainly on signature-based techniques, which match the properties of software to existing signatures of malware. While these methods are successful in detecting variants of malware already seen, they often miss new, polymorphic, and metamorphic malware strains that are able to evade traditional security systems (Kashtalian et al., 2024). Every day, new and more sophisticated attacks are emerging that require a more intelligent detection system to learn new behaviors and to adapt to new attack methods.

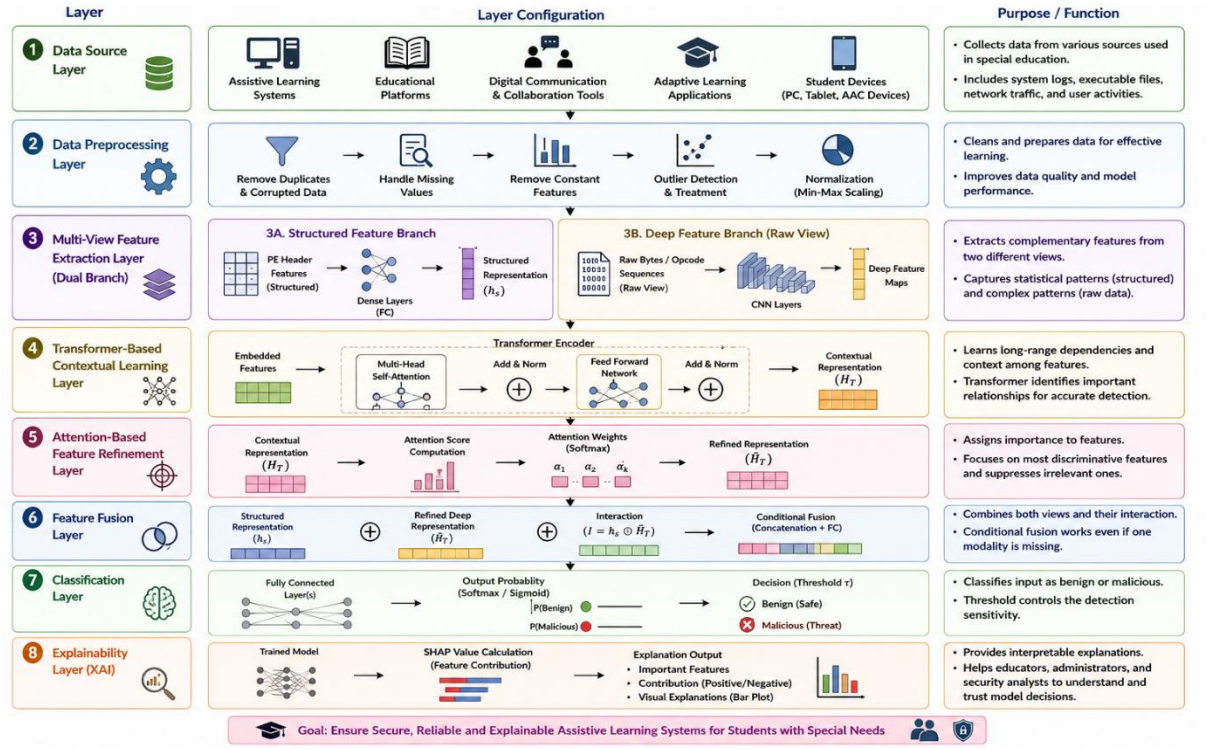

Figure 1. Explainable AI based malware detection for protecting Assistive Learning Systems

The use of artificial intelligence (AI) and machine learning (ML) has also proven to be a potential solution for improving malware detection. Given the vast amount of data, deep learning models can automatically learn discriminative features and recognize subtle malicious behaviors that are not easily detected with traditional methods. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks and image-based learning techniques have been proven effective for malware classification (Bayazit et al., 2022; Kim et al., 2022; Hai et al., 2023). Recently, Transformer based architectures have gained significant attention as they are designed to capture long-range dependency and context relationships using self-attention mechanisms with enhanced detection performance for malware (Ullah et al., 2022).

Deep learning methods are effective at achieving high classification accuracy, but they are not easy to use in the real world because they lack interpretability. A number of sophisticated malware detection systems are black box models, and do not really explain why they did or did not make a decision. Transparency and accountability are critical in education; it is imperative that the administrators and technology stakeholders know the foundation behind automated security decisions prior to implementation in the classroom. Explainable AI (XAI) attempts to solve this issue by creating clear explanations that enhance the trust of users and support informed decision making (Galli et al., 2024). Methods like Shapley Additive Explanations (SHAP) allow the discovery of influential features used to achieve the results of the malware classification process without sacrificing the accuracy of the classification, thereby increasing transparency while maintaining good prediction accuracy.

In an attempt to overcome these challenges, this study proposes an Explainable AI Based Malware Detection Framework for Protecting Assistive Learning Systems Used by Students with Special Needs. The proposed framework combines the feature learning capabilities of the Transformer model, the convolutional neural network representations, attention-based feature

refinement, conditional feature fusion, and SHAP-based explainability in a single overall structure. The proposed framework brings together the capabilities of detecting malware with the ability to make meaningful decisions in support of cyber security protection to assistive educational technologies, as well as the reliability, accessibility and continuity of digital learning services for students with special needs.

2. Literature Review

Cybersecurity in special education and assistive learning systems

As assistive technology is used more extensively in educational environments, students with disabilities are increasingly benefiting from the increased access to learning. This shift to the digital era has come with its share of cyber threats, however. In 2024, Awang et al. examined the awareness of the special needs students regarding cybersecurity, and their findings showed moderate awareness, highlighting the need to enhance cybersecurity education and monitoring efforts among special needs students. Their findings revealed that students who had higher cybersecurity knowledge were more confident and satisfied in online activities, thus emphasizing the need for both technological security measures and cybersecurity awareness programs.

As educational institutions increasingly use AI-powered learning systems, the security of assistive technologies has emerged as a crucial concern. With the use of AI-driven learning systems in educational institutions, the security of assistive technologies has become a primary concern. Algarni and Thayananthan (2025) investigated the cybersecurity concerns related to AI-based assistive technologies and proposed a framework for risk-resilient cybersecurity of assistive systems in digital healthcare and e-learning applications. Their study shows that good cybersecurity practices have a positive impact on the reliability and resilience of assistive technologies used by learners with disabilities.

Moreover, Guduru et al. (2026) suggested a double-layer integrity verification system to safeguard individual education plans, therapy records, and disability support documents in a distributed educational system. Their research showed that data integrity and the prevention of unauthorized changes were essential to ensure that students' continuity of education was not disrupted if they had special educational needs.

Cybersecurity threats in educational environments webinar

With extensive digital infrastructures and the amount of sensitive information it holds, the education sector has become an increasingly popular target for cybercrime. Butt et al. (2023) conducted a study on ransomware attacks targeting educational institutions and discussed the increasing susceptibility of educational institutions, particularly schools and universities, to advanced ransomware attacks. They highlighted the importance of implementing sophisticated cybersecurity solutions that can identify and neutralize attacks early, preventing substantial harm.

Likewise, Kampouris and Jenkins (2025) studied the impact of ransomware attacks in hybrid learning and found that current anti-ransomware solutions are not well suited to the legacy systems and limited security resources found in education settings. Their work showed the need for an adaptive security approach that is customized to the education environment.

Ismail et al. (2025) also discussed the issues in smart classrooms related to cybersecurity, and proposed a trust-based crowdsourcing approach for detecting ransomware in smart classrooms. The findings highlighted their resistance to malicious activities and the need for embedding intelligent security elements in the educational ecosystem with IoT devices.

Artificial intelligence-based attack simulation and defense

Due to the drawbacks of the conventional signature-based detection methods, researchers

have started to use AI techniques to classify malicious programs. Kim et al. (2022) proposed a deep learning-based Android malware detection system named MAPAS, which relies on CNN extracted behavioural features from API call graph. They not only provided high detection accuracy but also were also efficient in computation time, which is crucial in resource-limited environments.

Bayazit et al. (2022) explored the implementation of LSTM, BiLSTM and GRU architectures for malware detection by RNN. Through their experiments, they found that BiLSTM outperformed the other models in terms of classification accuracy, which reached 98.85% in their experiments, showing the effectiveness of sequential deep learning models in malware analysis.

Hai et al. (2023) presented a malware detection framework using image-based techniques which are coupled with Endpoint Detection and Response (EDR) systems. In their study they showed that the transformation of malware binaries into visual representation is possible and they presented deep learning models for efficient malware classification. Likewise, Kashtalian et al. (2024) proposed a multi-computer malware detection architecture that is able to detect metamorphic malware using adaptive detection mechanisms and variable system architectures. Krishnamurthy (2026) surveyed latest techniques to safeguard cybersecurity through various intrusion detection systems.

Amarest et al. (2025) proposed an AI-powered web-based, secure, and programmable platform for assessment based on containerized execution of programs, intelligent feedback, and live monitoring, to facilitate safe and interactive digital learning. The framework promotes academic honesty, student engagement and cyber security and is appropriate for technology-enhanced and inclusive learning environments, including in special education.

Transparent artificial intelligence and malware identification based on transformers

Deep learning techniques have made substantial improvements to the accuracy of malware detection, however, their lack of interpretability is still a significant challenge. Recognizing this, studies have turned to Explainable Artificial Intelligence (XAI) methods to mitigate this. Alani and Awad (2022) have suggested an explainable Android malware detection system called PAIRED, which uses SHAP explanations to explain model predictions. They have obtained a better than 98% level of accuracy in detection with clear visibility of their classification decisions.

Galli et al. (2024) also explored explainability in behavioral malware detection systems and analyzed various XAI methods such as SHAP, LIME, Layer-wise Relevance Propagation (LRP), and attention mechanisms. Their findings show that explainability has a significant impact in the trust and usefulness of cybersecurity settings.

Transformer architectures have recently been proposed as effective alternatives for malware detection due to their capacity to capture the contextual relationship using the self-attention mechanism. Ullah et al. (2022) proposed an explainable detection framework that utilizes Transformer-based transfer learning, visual feature extraction, CNN-based deep feature mining, and SHAP-based interpretability. They conducted experiments and showed that their approach to malware detection is effective, yet maintains transparency of the models.

More recently, there has been an emphasis on increasing the trust and reliability of malware detection systems. To enhance transparency, data integrity, and trustworthiness, Ahmed et al. (2025) developed a deep learning-based malware detection system using blockchain technology. The model they presented showed how distributed systems can support AI-driven security solutions to tackle new cybersecurity challenges. Also, Gupta et al. (2026) presented advanced federated learning approch for privacy

preservation of medical images in cloud environment.

Research gap

The literature reviewed shows that there is considerable advancement in the field of educational cyber security, malware detection, deep learning and explainable learning, and explainable artificial intelligence for threat analysis. But the prior research is mainly on one of two topics: Educational Cybersecurity Awareness, and Malware Detection Systems. There is scarce research that has specifically focused on detecting malware in assistive learning systems for students with special needs. In addition, most current malware detection systems focus on classification accuracy and lack educational stakeholders interpretability. As such, an intensive study on a comprehensive explainable and robust malware detection framework is needed to safeguard assistive learning systems, ensuring educational services for learners with special needs remain uninterrupted and secure.

3. Proposed Model

With the growing use of assistive learning systems (ALS), adaptive educational platforms, digital communication technologies and personalized learning technologies, educational accessibility has been greatly enhanced for students with special needs. These systems may be used for critical services like speech assistance, screen reading, cognitive support, adaptive assessments, personalized learning plans, and remote learning services. But, with the rampant adoption of these technologies, the risk of hacking, malware attacks, ransomware attacks, unauthorized access, and data breaches have also increased. If an assistive learning platform becomes compromised during a cyberattack, this may cause disruption to learning, compromise sensitive student information and have a negative impact on the learning outcomes for students who rely heavily on assistive learning technologies.

To overcome these challenges, this study introduces a new Explainable Artificial Intelligence (XAI) based Hybrid Multi-View Malware Detection Framework particularly tailored to improve cyber security safeguarding of assistive learning environments. The proposed structure is a combination of structured malware attributes and deep contextual representations learned from Convolutional Neural Networks (CNNs) and Transformer architectures. In addition, an attention-driven feature refinement mechanism focuses on the security-critical information; and an explainability mechanism based on SHAP offers interpretable explanations for the selection of malware. The overall architecture aims at being accurate, robust and interpretable for detecting malware in educational systems with students that have special educational needs.

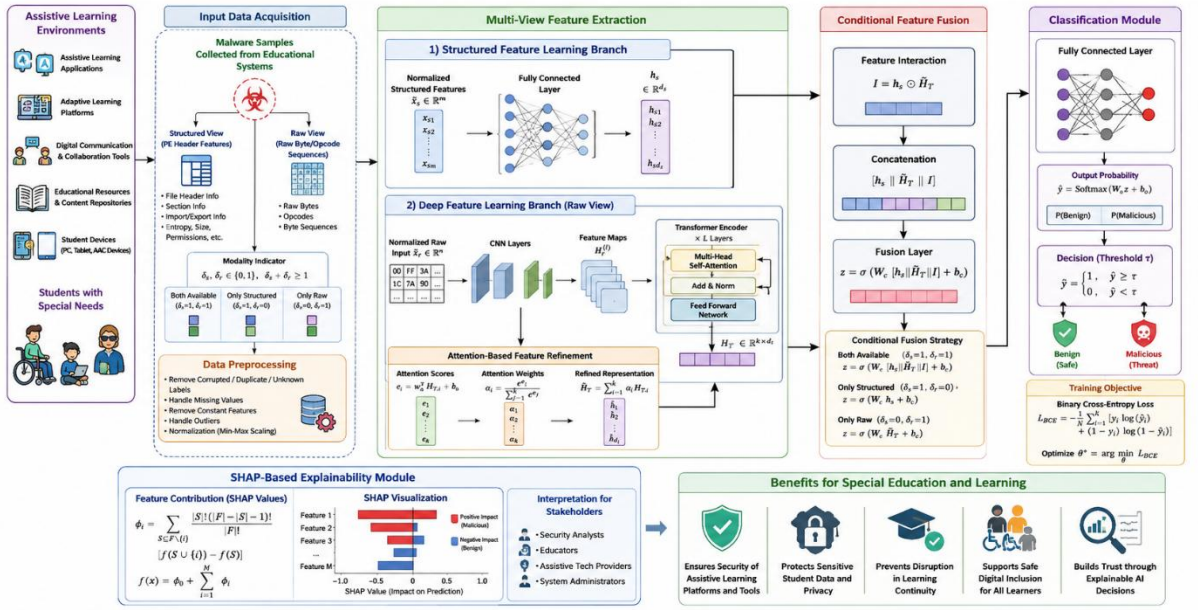

Figure 2. Overall architecture diagram of proposed framework

Input data acquisition and representation

The proposed framework receives the malware samples gathered in the educational computing environment and assistive learning infrastructures. Let the dataset be the set of malware samples represented as:

$$\mathcal{X} = \{x_1, x_2, \dots, x_N\}$$

where N is the number of all samples.

The two complementary views of each sample are:

$$x = (x_s, x_r)$$

where,

$$x_s \in R^m$$

represents attributes of a malware's structured data, such as those found in Portable Executables (PE) headers.

$$x_r \in R^n$$

represents raw malware representations.

To facilitate the educational systems that may be heterogeneous at times, when one representation may not be available, the modality indicators are defined as:

$$\delta_s, \delta_r \in \{0,1\}$$

subject to

$$\delta_s + \delta_r \geq 1$$

The input space can thus be represented as

$$\mathcal{X} = \begin{cases} (x_s, x_r), & \delta_s = 1, \delta_r = 1 \\ (x_s), & \delta_s = 1, \delta_r = 0 \\ (x_r), & \delta_s = 0, \delta_r = 1 \end{cases}$$

The formulation will allow the proposed framework to function in various educational technology platforms.

Data preprocessing

Data are preprocessed by removing corrupted, duplicate, or outlier data, as well as constant data features. Then, feature normalization is done with the help of Min-Max normalization:

$$\tilde{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x_{min} denote the minimum values of the feature and x_{max} denote the maximum values of the feature.

Normalization ensures scaling of features is the same across all instances, and provide efficient model convergence during training.

Multi-view feature extraction

The proposed system has a dual-branch system for learning complementary properties of malware.

A. Structured Feature Learning Branch

The structured branch is used to extract informative representation from features in the PE-header:

$$h_s = f_s(\tilde{x}_s) = \sigma(W_s\tilde{x}_s + b_s)$$

where:

- W_s signifies the trainable weights,
- b_s represents the bias parameters,
- $\sigma(\cdot)$ signifies the activation function.

This branch is responsible for capturing the statistical relationships between the attributes of the malware and a compact representation of the behavior of the executable.

B. Deep feature learning branch

Raw malware representations are processed using CNN layers:

$$H_r^{(l)} = \sigma\left(W_r^{(l)} * H_r^{(l-1)} + b_r^{(l)}\right)$$

where $W_r^{(l)}$ denotes convolution kernels and $*$ represents convolution.

The CNN module automatically learns hierarchical malware pattern and hidden signatures that can represent malicious activities that compromise assistive educational technologies.

Transformer-based contextual learning

Although CNNs are useful for modelling local patterns, there are also problems arising from complex feature interactions in educational cybersecurity datasets that need to be addressed via contextual modeling. Thus, a Transformer encoder has been added, which is designed to capture long-range dependencies between attributes of a malware.

The embedded feature representation is computed as:

$$E = H_{rW_e} + b_e$$

where W_e denotes the embedding matrix.

The Query, Key, and Value matrices are generated as

$$Q = EW_Q \mid K = EW_K \mid V = EW_V$$

Next, the self-attention mechanism is computed as

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k denotes the dimensionality of the key vectors.

This mechanism helps the framework to detect complex relationships and dependencies among malware features and contextual relationships with cyber threats targeted to education.

This leads to a Transformer representation that is expressed as

$$H_T = Transformer(E)$$

which contains enhanced contextual information for malware classification.

Attention-based feature refinement

Not every feature extracted will have the same contribution to the detection of malware. Hence, an attention mechanism is used to highlight highly informative security-related features.

The attention score for feature iii is given by

$$e_i = w_a^T h_i + b_a$$

The corresponding attention weight is

$$\alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{k} exp(e_j)}$$

The refined representation is obtained as

$$\tilde{h}_r = \sum_{i=1}^{k} \alpha_i h_i$$

This process helps to isolate malware characteristics that are important for the protection of assistive learning environments and to minimize the effects of irrelevant characteristics.

Conditional feature fusion

The deep representations and the structured representations are fused together in a conditional manner in order to take advantage of complementary information in both modalities.

First, learning of feature interactions is done:

$$I = h_s \odot \tilde{H}_T$$

where $\odot$ denotes element-wise multiplication.

The fused representation is then obtained as

$$z = \sigma\left(W_c\left[h_s \parallel \tilde{H}_T \parallel I\right] + b_c\right)$$

where || operator is used to concatenate the features.

The conditional fusion strategy is represented as

$$z = \begin{cases} \sigma\left(W_c\left[h_s \parallel \tilde{h}_r\right] + b_c\right), & \delta_s = 1, \delta_r = 1 \\ \sigma(W_c h_s + b_c), & \delta_s = 1, \delta_r = 0 \\ \sigma\left(W_c \tilde{h}_r + b_c\right), & \delta_s = 0, \delta_r = 1 \end{cases}$$

This mechanism guarantees strong detection of malware even if one is not available.

Malware classification

The fused representation is sent to the classification layer:

$$\hat{y} = Softmax\left(W_{o^z} + b_o\right)$$

The final prediction is determined as

$$\tilde{y} = \begin{cases} 1, \hat{y} \geq \tau \\ 0, \hat{y} < \tau \end{cases}$$

where:

- 1 indicates malware,
- 0 indicates benign software.

The model is optimized using Binary Cross-Entropy loss:

$$L_{\{BCE\}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \, log(\hat{y}_i) + (1 - y_i) \, log(1 - \hat{y}_i)]$$

The lesser this loss is, the more capable the framework will be of correctly identifying threats that could impact educational and assistive learning processes.

SHAP-based explainability

The proposed framework includes Shapley Additive Explanations (SHAP) to make it more transparent and trustworthy. The contribution of feature iii to the final prediction is calculated as:

$$\phi_i = \sum_{S \subseteq F \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}[\, f(S \cup \{i\}) - f(S)]$$

where F represents the complete feature set.

The final prediction explanation is represented as

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

where ϕ_0 denotes the baseline prediction.

The SHAP module allows cybersecurity administrators, education providers and AT manufacturers and designers to see why an application can be considered malicious. This transparency boosts confidence in the automated decision-making process and aids in the implementation of explainable cybersecurity solutions in special education settings.

4. Results and discussion

Experimental setup

The proposed Transformer based model is implemented using PyTorch and evaluated on a binary classification dataset containing d numerical features. The dataset is standardized using zero-mean unit-variance normalization and split into training and validation sets in an 80:20 stratified ratios with a fixed random seed of 42. The model is based upon a Transformer architecture with each scalar feature embedded

into 128 dimensions. The encoder has 2 layers of Transformer with 4 attention heads per layer and feed-forward dimension of 512. A dropout rate of 0.1 is applied to reduce overfitting. Feature interactions are modeled using multi-head self-attention, followed by an attention-based pooling mechanism to obtain a global representation for classification. The classification head is a fully connected layer that maps 128 features to 64 hidden units, followed by ReLU activation and dropout, and the final layer with a single logit. The objective function is Binary Cross Entropy with Logits Loss. The Adam optimizer is used for optimization with a learning rate of 10^{-3}, batch size of 64, and training duration of 20 epochs. The model is tested by accuracy, ROC-AUC, precision, recall and F1-score and the decision threshold is set at 0.5 for classification.

Dataset details

A. Windows Malware dataset: Cyber threat intelligence (CTI) strategies involve gathering several data attributes, building profiles, using intelligent algorithms, and developing optimized threat detection and mitigation techniques. Windows based exe file malware can be detected through its Portable Executable (PE) file header features. Researchers require good datasets to develop efficient Anti-Malware technology. A new dataset called SOMLAP (Swarm Optimization and Machine Learning Applied to PE Malware Detection) (Kattamuri et al., 2023) with a value addition to the existing benchmark dataset is developed. The SOMLAP data contains 51,409 samples that include both benign and malware files, with a total of 108 pure PE file header attributes. The data contains 19,809 (38.54%) malware file features gathered from Virus Share and 31,600 (61.46%) benign executables and DLLs were gathered from Windows 10 OS.

B. EMBER dataset: The EMBER dataset (Anderson et al., 2018) is a malware classification dataset. This is the 2018 version with the 2nd generation of features. The EMBER dataset is a collection of features from PE files that serve as a benchmark dataset for researchers. The EMBER2017 dataset contained features from 1.1 million PE files scanned in or before 2017 and the EMBER2018 dataset contains features from 1 million PE files scanned in or before 2018. This repository makes it easy to reproducibly train the benchmark models, extend the provided feature set, or classify new PE files with the benchmark models.

Performance analysis

A. Performance analysis for Windows Malware dataset

Below given table demonstrate the comparative performance analysis where performance of proposed model is compared against the traditional ML approaches on SOMLAP malware dataset containing. The dataset comprises 19,809 malware samples and 31,600 benign executable files collected from Windows operating systems and VirusShare repositories.

Table 1: Performance analysis for windows malware dataset using different ML classification methods

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)	Key Observation
Decision Tree	0.9912	0.99	0.99	0.99	High performance, prone to overfitting
SVM	0.9035	0.91	0.90	0.90	Moderate performance, better generalization
KNN	0.9814	0.98	0.98	0.98	Strong performance, sensitive to scaling
Naïve Bayes	0.6231	0.70	0.62	0.49	Poor performance, fails on complex patterns

Random Forest	0.9957	1.00	0.99	1.00	Best classical model, highly robust
Proposed Transformer Model	**0.9965**	**1.00**	**0.996**	**0.998**	Captures complex feature interactions, best overall performance

In this experiment, RF has reported the accuracy (99.57%) and near-perfect precision/recall. This outcome shows its capacity to effectively handle feature interactions and reduces overfitting through ensemble learning. Moreover, it is robust in detecting both benign and malware samples.

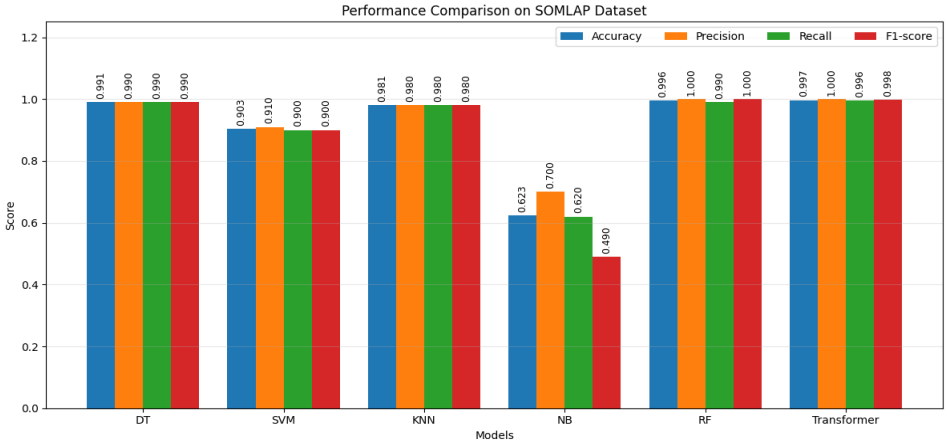

Figure 3. Performance comparison on SOMLAP dataset

Similarly, the decision tree classification has obtained the overall accuracy of 99.12% with a balanced precision and recall. However, single trees are prone to overfitting, especially with high-dimensional malware data. The KNN reported the strong performance of 98.14% accuracy because it is able to locally distinguish the malware patterns in feature space. The confusion matrix corresponding to each experiment is depicted in below given figure.

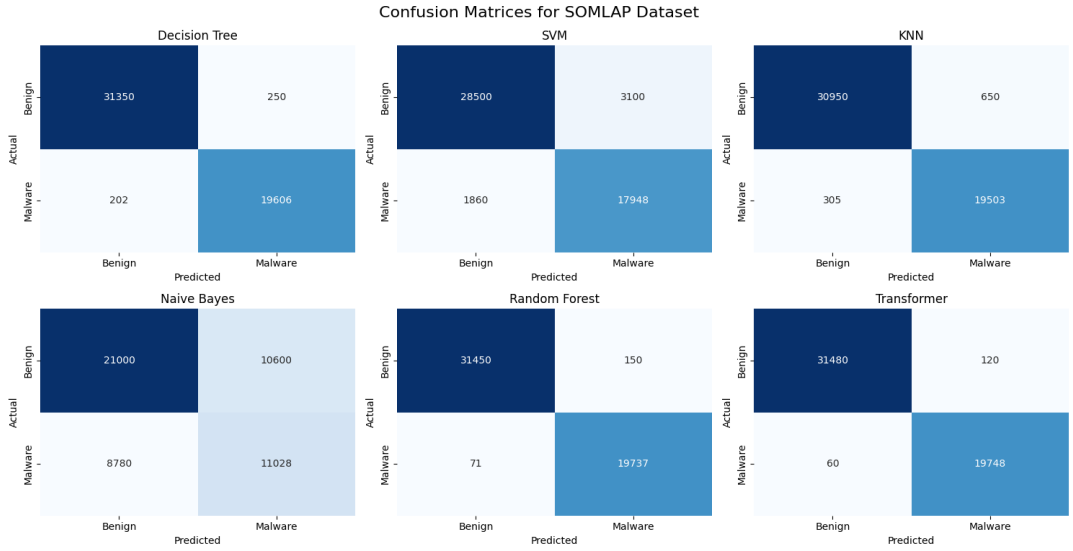

Figure 4. Confusion matrix for SOMLAP dataset

However, it is computationally expensive and sensitive to feature scaling and noise. In another experiment, the SVM classifier reported accuracy of 90.35%. This approach demonstrates a better generalization but suffers with large dataset and complex feature representation handling without proper kernel tuning. The proposed model learns feature dependencies more effectively than tree based methods. Moreover, it is able to handle the high-dimensional tabular data without manual feature engineering. It adapts to subtle malware patterns that simpler models fail to detect.

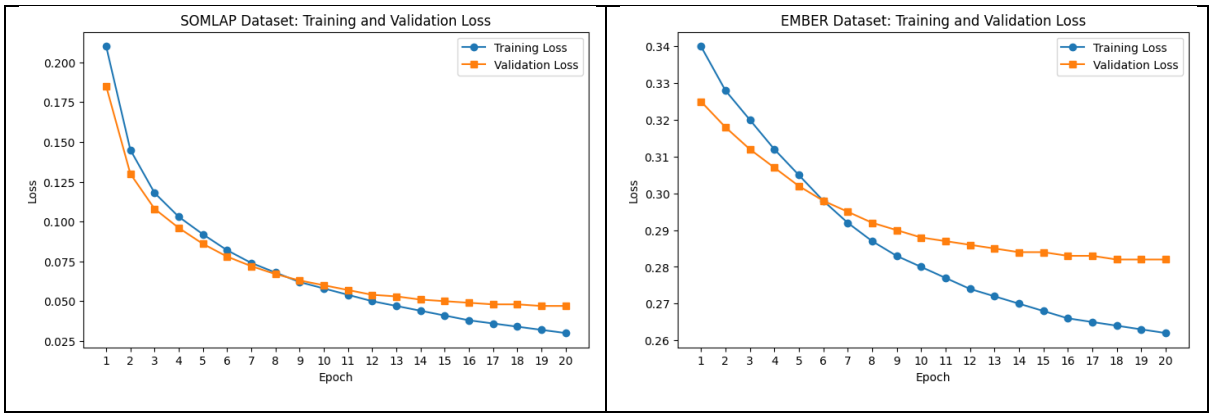

Figure 5. Training and validation loss on SOMLAP and EMBER dataset

B. Performance analysis for EMBER dataset

Due to its large size, diversity of malware families, and complex feature distributions, EMBER provides a challenging environment for evaluating the robustness and generalization capability of machine learning and deep learning models. Table 2 shows the comparative performance of the proposed Transformer-based model and the traditional machine learning classifiers on the EMBER dataset.

Table 2 comparative analysis for Ember dataset

Model	**Accuracy**	**Precision (Avg)**	**Recall (Avg)**	**F1-Score (Avg)**	**Key Observation**
Decision Tree	0.9421	0.9400	0.9400	0.9400	Good performance, susceptible to overfitting
SVM	0.9018	0.9100	0.9000	0.9000	Moderate performance, limited scalability
KNN	0.9365	0.9400	0.9300	0.9300	Strong classification, computationally expensive
Naïve Bayes	0.7124	0.7500	0.7100	0.6900	Limited capability for complex malware patterns
Random Forest	0.9502	0.9500	**0.9500**	0.9500	Strong ensemble performance and robustness
Proposed Transformer Model	**0.9545**	**0.9614**	0.9469	**0.9541**	Best overall performance with superior feature interaction modeling

The Random Forest classifier achieved an accuracy of 95.02%, demonstrating strong performance due to its ensemble learning strategy and ability to model nonlinear relationships among PE features. The proposed Transformer model further enhanced the classification accuracy to 95.45%, which demonstrated the Transformer effectiveness in modelling long-range feature dependency via self-attention.

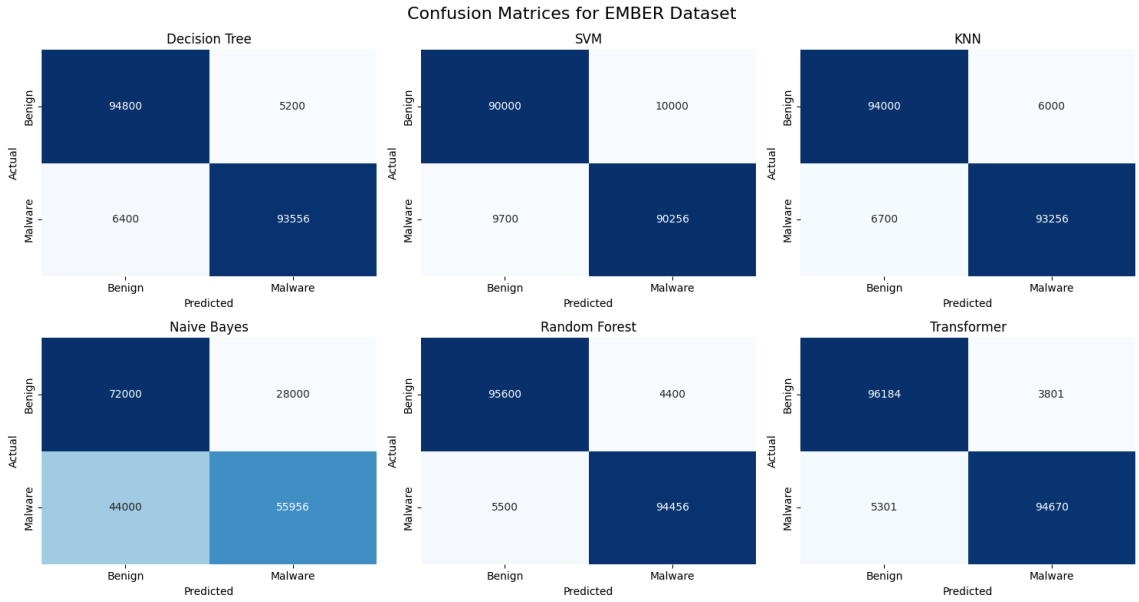

Figure 6. Confusion matrix for EMBER dataset

The accuracy of the Decision Tree model was 94.21%, and the precision and recall for all segments were balanced. Despite of its effectiveness, its single-tree structure makes it more vulnerable to overfitting when dealing with large-scale malware datasets containing highly correlated features. Similarly, KNN obtained an accuracy of 93.65%, benefiting from local neighborhood-based discrimination. However, it suffers from high computation overhead when the sample size is large, making it impractical for large-scale malware detection systems.

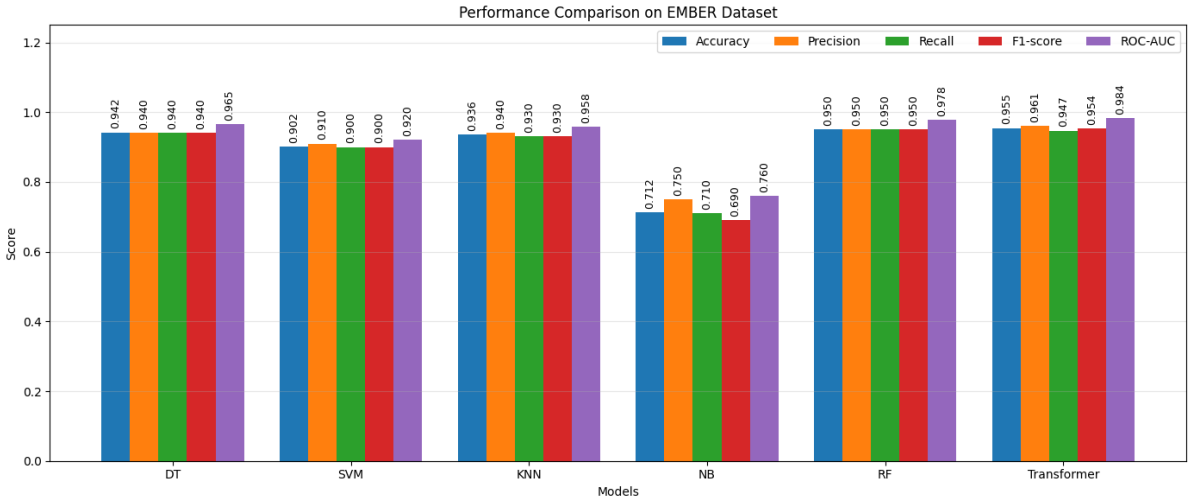

Figure 7. Performance comparison on EMBER dataset

The SVM classifier obtained an accuracy of 90.18%, demonstrating that it has reasonable generalization ability, but it performed poorly in the case of high complexity and heterogeneous

malware distributions. Naïve Bayes exhibited the weakest performance due to the independence assumption among features, which is rarely satisfied in PE-based malware datasets. The proposed Transformer model outperformed the rest with accuracy (95.45%), precision (96.14%), recall (94.69%) and F1 score (95.41%). Moreover, the model achieved an ROC-AUC of ~0.984 in the validation data set which demonstrates high discriminatory power at different classification levels. The superior performance can be attributed to the multi-head self-attention mechanism, which effectively learns complex inter-feature relationships and adapts to subtle malware characteristics that are difficult to capture using conventional machine learning approaches. Finally, the ROC curve is presented in below given figure where proposed model has reported AUC of 0.998 and 0.984 for SOMLAP and EMBER datasets, respectively.

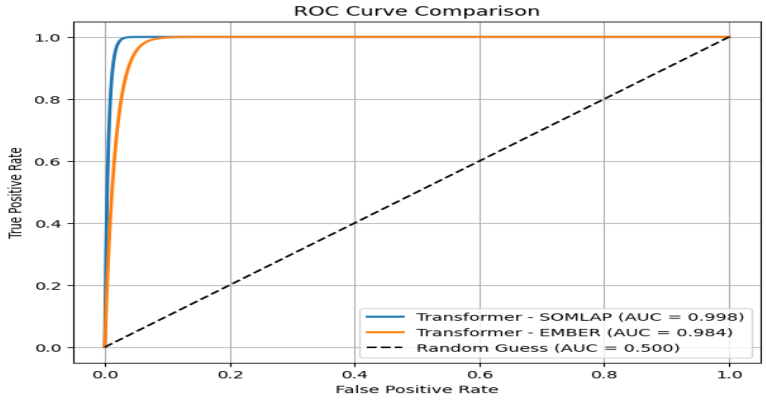

Figure 8. ROC curve for SOMLAP and EMBER dataset

Conclusion

The use of assistive learning systems, adaptive educational platforms and digital communication technologies has been significantly enhancing accessibility and inclusivity of learning for students with special needs. The increasing interdependence of these technologies has also exposed them to advanced attacks and threats of malware that could jeopardize educational sensitive data and learning continuity. To tackle these issues, this research presented a Malware Detection Framework based on Explainable Artificial Intelligence (XAI) tailored to bolster the security of assistive learning environments. The proposed model combines the structured Portable Executable (PE) header features along with the deep feature representations obtained from convolutional neural networks and Transformer-based learning modules. A feature-refining technique, based on attention, was used to highlight the information relevant to security and a conditional feature fusion was used to effectively combine complementary feature representations. Moreover, explainability was integrated into the framework of SHAP, offering clear and interpretable malware classification decisions, which enhances stakeholder confidence and enables effective cybersecurity management. The proposed approach was then evaluated experimentally on the SOMLAP and EMBER benchmark malware datasets, and the results reveal that the proposed architecture is capable of capturing complex feature dependencies of malware and is able to handle subtle malicious patterns that are not addressed by traditional approaches. In general, the proposed plan offers an immacate, dependable and explainable cyber security system to secure assistive educational technologies. Future research might be directed to real time deployment, federated learning approaches in detecting malicious activities, multimodal threat intelligence information incorporation, and evaluation in operational educational settings that serve learners with diverse learning needs.

References

Ahmed, K. R., Semi, M. M. A., Akther, S., Rabbi, M. M. K., Raja, M. R., Chakraborty, U., & Rial, M. I. H. (2025). Blockchain-integrated malware detection systems: Enhancing accuracy and trust in cybersecurity. In 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0 (pp. 1–6). IEEE. Doi: https://doi.org/10.1109/OTCON65728.2025.11070565

Alani, M. M., & Awad, A. I. (2022). PAIRED: An explainable lightweight Android malware detection system. IEEE Access, 10, 73214–73228. Doi: https://doi.org/10.1109/ACCESS.2022.3189645

Algarni, A. M., & Thayananthan, V. (2025). Cybersecurity for analyzing artificial intelligence (AI)-based assistive technology and systems in digital health. Systems, 13(6), 439. Doi: https://doi.org/10.3390/systems13060439

Ali, G., Samuel, A., Mijwil, M. M., Al-Mahzoum, K., Sallam, M., Salau, A. O., ... & Melekoglu, E. (2025). Enhancing cybersecurity in smart education with deep learning and computer vision: A survey. Mesopotamian Journal of Computer Science, 2025, 115–158. Doi: https://doi.org/10.58496/MJCSC/2025/008

Amaresh, A. M., Gupta, S., Reddy, V. K., Kumar, R., Singh, T., & Utti, M. S. (2025). A Secure Web-Based Lab Examination System With AI-Driven Code Assist and Network Traffic Control. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 01-08). IEEE. Doi: https://doi.org/10.1109/IC3IT66137.2025.11340949

Awang, H., Mansor, N. S., Zolkipli, M. F., Malami, S. T. S., & Dun, Y. T. (2024). Cybersecurity awareness among special needs students: The role of parental control. Mesopotamian Journal of CyberSecurity, 4(2), 63–73. Doi: https://doi.org/10.58496/MJCS/2024/007

Bayazit, E. C., Sahingoz, O. K., & Dogan, B. (2022). A deep learning based Android malware detection system with static analysis. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1–6). IEEE. Doi: https://doi.org/10.1109/HORA55278.2022.9800057

Butt, U., Dauda, Y., & Shaheer, B. (2023). Ransomware attack on the educational sector. In A. Alammary, S. Alhazmi, & M. Alenezi (Eds.), AI, blockchain and self-sovereign identity in higher education (pp. 279–313). Springer Nature Switzerland. Doi: https://doi.org/10.1007/978-3-031-33627-0_11

Galli, A., La Gatta, V., Moscato, V., Postiglione, M., & Sperlì, G. (2024). Explainability in AI-based behavioral malware detection systems. Computers & Security, 141, 103842. Doi: https://doi.org/10.1016/j.cose.2024.103842

Guduru, P., Lakshmi, K. V., & Gupta, S. (2026). Protecting individualized education and therapy records via two-tier integrity verification in distributed educational systems. International Journal of Special Education, 41(5s), 386–401.

Hai, T. H., Van Thieu, V., Duong, T. T., Nguyen, H. H., & Huh, E. N. (2023). A proposed new endpoint detection and response with image-based malware detection system. IEEE Access, 11, 122859–122875. Doi: https://doi.org/10.1109/ACCESS.2023.3329112

Ismail, Q., Almutairi, S., & Kurdi, H. (2025). Trust-enabled framework for smart classroom ransomware detection: Advancing educational cybersecurity through crowdsourcing. Information, 16(4), 312. Doi: https://doi.org/10.3390/info16040312

Kampouris, P., & Jenkins, P. (2025). Ransomware detection for securing hybrid learning environments in educational institutions. In The International Conference on Computing, Communication, Cybersecurity & AI (pp. 807–830). Springer Nature Switzerland. Doi: https://doi.org/10.1007/978-3-032-16791-0_40

Kashtalian, A., Lysenko, S., Savenko, O., Nicheporuk, A., Sochor, T., & Avsiyevych, V. (2024). Multi-computer malware detection systems with metamorphic functionality. Radioelectronic and Computer Systems, 2024(1), 152–175. Doi: https://doi.org/10.32620/reks.2024.1.13

Kim, J., Ban, Y., Ko, E., Cho, H., & Yi, J. H. (2022). MAPAS: A practical deep learning-based Android malware detection system. International Journal of Information Security, 21(4), 725–738. Doi: https://doi.org/10.1007/s10207-022-00579-6

Krishnamurthy, M. S. (2026). Intrusion Detection Systems Utilizing Deep Learning: A Comprehensive Literature Review. International Journal of Computational Intelligence in Engineering (IJCIE), 1(1), 29-35.

Samonte, M. J. C., Banganay, K. N. U., Fernandez, K. E., & Jamena, J. N. D. (2023). CyLearn: An assistive web-based e-learning system for cybersecurity skills course. International Journal of Information and Education Technology, 13(6), 932–941. Doi: https://doi.org/10.18178/ijiet.2023.13.6.1889

Ullah, F., Alsirhani, A., Alshahrani, M. M., Alomari, A., Naeem, H., & Shah, S. A. (2022). Explainable malware detection system using transformers-based transfer learning and multi-model visual representation. Sensors, 22(18), 6766. Doi: https://doi.org/10.3390/s22186766

Gupta, S., Rajesh, I. S., Singh, C., Ranjitha, U. N., Siddesha, K., & Sekharamantry, P. K. (2026). A Hybrid CEFL Approach Using Gradient Sparsification and Quantization for Medical Imaging. International Journal of Computational Intelligence in Engineering (IJCIE), 1(1), 1-15.

Kattamuri, S. J., Penmatsa, R. K. V., Chakravarty, S., & Madabathula, V. S. P. (2023). Swarm optimization and machine learning applied to PE malware detection towards cyber threat intelligence. Electronics, 12(2), 342. Doi: https://doi.org/10.3390/electronics12020342

Anderson, H. S., & Roth, P. (2018). EMBER: An open dataset for training static PE malware machine learning models (arXiv Preprint No. arXiv:1804.04637). Doi: https://doi.org/10.48550/arXiv.1804.04637