Sr. No	Title	Authors	Publication name & year	Proposed Methodology	Results	Conclusion/ Limitations
1	Detection of malware in downloaded files using various machine learning models	Akshit Kamboj, Priyanshu Kumar, Amit Kumar Bairwa, Sandeep Joshi	Egyptian Informatics Journal Volume 24, Issue 1, March 2023	The proposed methodology aims to leverage machine learning techniques for the detection of malware in downloaded files. By analyzing features such as MD5 hash, Optional Header size, and Load Configuration Size, various classification algorithms including Decision Tree, Random Forest, Gradient Boosting, XGBoost, and AdaBoost will be trained and evaluated.	The results that have been observed in the case of supervised machine learning models are much better as compared to those obtained in the case of unsupervised machine learning models. Among supervised machine learning models, the Random Forest Model is the most accurate. On the other hand, in the case of unsupervised machine learning models, Principal Component Analysis has been observed to achieve the highest accuracy.	The project used supervised (e.g., Logistic Regression, Decision Tree, Random Forest) and unsupervised (e.g., PCA, K-Means) machine learning models on a dataset, employing balancing techniques. Random Forest performed best with 99.99% accuracy, followed by XGBoost (99.68%), Decision Tree (99.57%), Gradient Boosting (99.09%), and AdaBoost (98.87%). Compared to previous research, four out of five models achieved higher accuracies.
2	Automated, Reliable Zero-Day Malware Detection Based on Autoencoding Architecture	C. Kim, S. -Y. Chang, J. Kim, D. Lee and J. Kim	*IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 3900-3914, Sept. 2023, doi: 10.1109/TNSM.2023.3251282	Current malware detection methods often rely on known patterns, limiting their effectiveness against zero-day threats. Semi-supervised learning can address this but faces challenges like sensitivity to thresholds in autoencoder-based approaches and low detection rates in one-class classification. Our method combines autoencoding and one-class classification to overcome these limitations, achieving high detection rates for both known and unknown malware patterns.	Demonstrating the effectiveness of the proposed method in detecting both known and unknown malware patterns. The method combines autoencoding and one-class classification, achieving high detection rates with up to 97.1% accuracy on public malware datasets. Additionally, the approach shows resilience to evasion attacks, maintaining over 99% detection rates.	In conclusion, the paper introduces an innovative approach for automated, reliable zero-day malware detection based on an autoencoding architecture, addressing critical limitations of existing methods and demonstrating high accuracy rates on public malware datasets while also showing resilience against evasion attacks.

Sr. No	Title	Authors	Publication name & year	Proposed Methodology	Results	Conclusion/ Limitations
3	A Comprehensive Review on Malware Detection Approaches.	Ö. A. Aslan and R. Samet	*IEEE Access*, vol. 8, pp. 6249-6271, 2020, doi: 10.1109/ACCESS.2019.2963724	conducting a thorough review of existing malware detection approaches. This includes examining various techniques such as signature-based detection, behavior-based detection, machine learning-based approaches, and hybrid methods. The review aims to analyze the strengths and weaknesses of each approach, identify emerging trends, and provide insights into the current state of malware detection technology	The research provides an extensive overview of existing methods for detecting malware. It covers various techniques including signature-based detection, behavior-based detection, machine learning approaches, and hybrid methods. The review aims to evaluate the effectiveness of these approaches, highlight their strengths and weaknesses, and identify emerging trends in the field of malware detection.	The paper discusses challenges in malware detection, noting strengths and limitations of existing methods. It highlights the need for innovation and suggests combining different detection approaches for better results. Overall, it's a valuable reference for enhancing cybersecurity efforts.
4	Machine Learning for Detecting Malware in PE Files	Connors, Collin,Sarkar, Dilip	Research Gate, 2022.	The proposed methodology likely involves using machine learning to analyze Portable Executable (PE) files for malware detection. This includes feature extraction from PE file attributes and training machine learning algorithms to classify files as benign or malicious based on these features.	The results of the paper indicate that artificial neural network (ANN)-based systems outperform other machine learning (ML)-based systems for malware detection. Additionally, the study showed that not all features contribute equally to the performance of the detection system. Specifically, the study found that Byte histogram and Byte entropy, although performing well individually, did not significantly improve the system's performance	In conclusion, the study highlights the effectiveness of artificial neural network (ANN)-based systems for malware detection using the EMBER dataset. It emphasizes that not all features contribute equally to detection performance and that the combination of certain features may not significantly improve system performance. Specifically, the study identifies the Byte histogram and Byte entropy features as an example of this

Sr. No	Title	Authors	Publication name & year	Proposed Methodology	Results	Conclusion/ Limitations
5	Analyzing and comparing the effectiveness of malware detection: A study of machine learning approaches	Muhammad Azeem, Danish Khan, Saman Iftikhar, Shaikhan Bawazeer, Mohammed Alzahrani	Heliyon, Volume 10, Issue 1, 2024	The proposed methodology likely involves selecting datasets, extracting features, preprocessing data, choosing machine learning models, training and evaluating these models, comparing their performance, and interpreting the results to understand the effectiveness of different machine learning approaches for malware detection.	The study proposes a technique for malware detection and classification, emphasizing preprocessing, feature extraction, and entropy-based feature selection. By dropping various features and employing feature encoding, the approach achieved high accuracy, notably 99.98% on the ET classifier. The study suggests further exploration of Hidden Markov Models and Deep Learning methods for enhanced prediction models.	While our proposed technique shows promising results, there are certain limitations to consider. The study's reliance on specific classifiers may restrict generalizability, and the dataset's size and composition may influence the effectiveness of the approach. Additionally, the use of feature encoding for categorical and numerical values may introduce biases. Further research should address these limitations and explore alternative methods to improve the robustness
6	Malware Analysis and Detection Using Machine Learning Algorithms	Akhtar, Muhammad & Feng, Tao	Research gate, 2022	The study employs various machine learning techniques to detect polymorphic malware. These include Naive Bayes, SVM, J48 decision trees, Random Forest, and a proposed approach. Detection accuracy is evaluated using high detection ratios and the confusion matrix. Results show that DT, CNN, and SVM algorithms perform well in detecting malware, with low false positive rates.	The study conducted training and testing phases for malware detection using classifiers such as KNN, CNN, NB, RF, SVM, and DT. Results showed DT with the highest accuracy (99%) and TPR (99.07%), making it the optimal model for malware detection. SVM and CNN followed closely, demonstrating comparable high accuracy rates. Specifically, using DT yielded the best performance for identifying malware, with significantly higher TPR and accuracy compared to other classifiers.	In conclusion, the paper showcases the rising interest in ML algorithms for malware detection. It evaluates three ML approaches and finds that Decision Trees (DT), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM) perform well, with DT achieving the highest accuracy (99%). The study emphasizes the potential of static analysis based on PE data and feature selection in accurately identifying malware without execution, offering a significant

Sr. No	Title	Authors	Publication name & year	Proposed Methodology	Results	Conclusion/ Limitations
7	Malware Detection Using Byte Streams of Different File Formats	Y. -S. Jeong, S. -M. Lee, J. -H. Kim, J. Woo and A. R. Kang	IEEE Access, 2022	This paper aims at solving the malware detection task that is basically a binary classification; we want to develop a model that predicts a label (malware or benign) of a given byte stream of a non-executable. The overview of our method is depicted in Fig. 1. We basically utilize byte streams of multiple file formats for training a classification model for a specific file format.	In our experiments, we evaluated malware detection performance for different formats using MalConv and SPAPConv models. Training with both HWP and PDF byte streams together improved precision for HWP malware detection, especially with MalConv. However, SPAPConv did not show performance improvement. Similarly, for PDF malware detection, MalConv showed slight performance enhancement, while SPAPConv did not improve. This difference could be due to model size and architecture variations.	Our study investigated using byte streams from various file formats for non-executable malware detection, revealing potential performance enhancements. We observed that combining byte streams from different formats may improve detection accuracy. Additionally, analysis showed common JavaScript functions in HWP and PDF formats, suggesting a potential reason for performance improvements with combined streams. In future research, we plan to collect and annotate more data to explore the impact of using different stream formats on various
8	Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions	M. Aljabri et al.	IEEE Access, vol. 10	The proposed methodology involves reviewing studies on detecting malicious URLs using ML and DL techniques from 2012 to 2021. The review categorizes research based on language, URL features, ML methods, and datasets. It includes comparisons and discussions on detection techniques and properties. Feature extraction focuses on lexical, content-based, and network-based features. Challenges like dataset size and feature selection are discussed.	A review of various machine learning algorithms for malicious URL detection found promising results, with Random Forest and SVMs exceeding 90% accuracy in some cases. However, challenges like class imbalance and concept drift necessitate further research. Future directions include exploring deep learning, online learning, and user feedback integration to continuously adapt models and improve real-time detection capabilities.	This review identified effective machine learning techniques for detecting malicious URLs in both Arabic and non-Arabic content. Lexical features dominate, while Arabic studies lack network-based features. SVM, RF, and NB are popular algorithms, but CNN and XGBoost achieve higher accuracy. Studies often use custom datasets for Arabic content, while open-source ones like PhishTank are common for non-Arabic.

Sr. No	Title	Authors	Publication name & year	Proposed Methodology	Results	Conclusion/ Limitations
9	Detecting malicious URLs using machine learning techniques	F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof and M. Köppen	2016 IEEE Symposium Series on Computational Intelligence (SSCI)	In this paper, the detection of malicious URLs as a binary classification problem and study the performance of several well-known classifiers, namely Naïve Bayes, Support Vector Machines, Multi-Layer Perceptron, Decision Trees, Random Forest and k-Nearest Neighbors.	This study evaluated various machine learning algorithms for detecting malicious URLs. Random Forest and Multi-Layer Perceptron (MLP) emerged as the top performers, achieving high accuracy, precision, and recall. Interestingly, a simple feature selection method using Pearson correlation yielded the best results, showcasing its effectiveness compared to more complex techniques. Furthermore, numerical features alone proved surprisingly adequate for training, suggesting their usefulness even without domain expertise. Overall, the study highlights the importance of feature selection and the potential of both Random Forest and MLP for accurate URL classification.	Overall, this study demonstrates the effectiveness of machine learning in detecting malicious URLs and emphasizes the importance of feature selection strategies. Further research across diverse datasets and incorporating domain knowledge could offer even more robust and accurate detection capabilities.