Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055463

ARTICLE

TLERAD: Transfer Learning for Enhanced Ransomware Attack Detection

Isha Sood* **and Varsha Sharma**

School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, 462033, India

*Corresponding Author: Isha Sood. Email: ishasweet1984@gmail.com

Received: 27 June 2024 Accepted: 05 September 2024 Published: 18 November 2024

ABSTRACT

Ransomware has emerged as a critical cybersecurity threat, characterized by its ability to encrypt user data or lock devices, demanding ransom for their release. Traditional ransomware detection methods face limitations due to their assumption of similar data distributions between training and testing phases, rendering them less effective against evolving ransomware families. This paper introduces TLERAD (Transfer Learning for Enhanced Ransomware Attack Detection), a novel approach that leverages unsupervised transfer learning and co-clustering techniques to bridge the gap between source and target domains, enabling robust detection of both known and unknown ransomware variants. The proposed method achieves high detection accuracy, with an AUC of 0.98 for known ransomware and 0.93 for unknown ransomware, significantly outperforming baseline methods. Comprehensive experiments demonstrate TLERAD's effectiveness in real-world scenarios, highlighting its adaptability to the rapidly evolving ransomware landscape. The paper also discusses future directions for enhancing TLERAD, including real-time adaptation, integration with lightweight and post-quantum cryptography, and the incorporation of explainable AI techniques.

KEYWORDS

Ransomware detection; transfer learning; unsupervised learning; co-clustering; cybersecurity; machine learning; lightweight cryptography; post-quantum cryptography; explainable AI; TLERAD

1 Introduction

Ransomware has rapidly become one of the most prevalent and damaging forms of cyberattack, characterised by its malicious intent to either encrypt user data or lock devices, rendering them inaccessible until a ransom is paid. These threats typically employ sophisticated encryption techniques or device lockout mechanisms, ultimately culminating in ransom demands. The impact of ransomware is wide-ranging, affecting individuals, corporations, and even government institutions, leading to substantial financial losses, operational disruptions, and compromised data integrity. Ransomware can be broadly categorized [1] into two types:

- **Locker Ransomware:** This type of ransomware restricts access to a computer's interface, typically leaving the core data intact. While the data remains unaltered, users are locked out of their systems, often by blocking access to the operating system or critical system utilities, making it impossible to use the device without paying the ransom.

- **Crypto Ransomware:** This variant is more insidious, as it encrypts critical user data, making the information inaccessible without a specific decryption key. The encryption used is often highly sophisticated, ensuring that brute force decryption is practically impossible within a reasonable timeframe. Victims are coerced into paying the ransom to retrieve their data, though payment does not always guarantee decryption.

Historically, ransomware has existed for decades, with its origins traced back to the late 1980s. However, the period between 2013 and 2014 marked a dramatic escalation in the threat landscape, with a 250% surge in the emergence of new crypto-ransomware families [2]. This surge can be attributed to several factors, including the proliferation of cryptocurrencies like Bitcoin, which provide attackers with a relatively anonymous and untraceable payment method. The WannaCry attack in 2017 further underscored the escalating threat landscape [3], infecting over 200,000 computers across 150 countries within a day. WannaCry exploited a vulnerability in the Windows operating system, leading to widespread disruption in critical sectors, including healthcare, finance, and transportation, and highlighting the urgent need for more effective countermeasures.

In response to these growing threats, cybersecurity researchers and professionals have increasingly turned to machine learning as a potent defense mechanism against ransomware and other forms of malware [4]. Machine learning algorithms excel at identifying patterns and anomalies within large datasets, making them well-suited for detecting malicious activities that deviate from normal behavior. These algorithms can be trained to recognise the signatures of known ransomware strains and to identify potentially harmful behaviors in previously unseen strains.

However, one of the significant challenges in applying machine learning to ransomware detection is the assumption that the probability distributions of training and test data are consistent over time [5]. This assumption often fails in real-world scenarios, where the rapid evolution of ransomware results in distribution shifts that traditional models struggle to handle. Ransomware developers continuously adapt their strategies, creating new variants that evade existing detection mechanisms. This evolving threat landscape necessitates the development of more adaptive models that can maintain high detection accuracy despite the variability in data distributions.

This study addresses this critical gap by enhancing ransomware detection through the integration of machine learning and transfer learning techniques. Transfer learning, a subfield of machine learning, involves transferring knowledge gained from one domain or task to improve performance on a related but different domain or task. In the context of ransomware detection, transfer learning can help bridge the gap between the varying distributions of training and test data, leading to more robust and adaptive detection models.

In particular, this study addresses this gap by enhancing ransomware detection through integrating machine learning and transfer learning. In particular we poposed a TLERAD: Transfer Learning for Enhanced Ransomware Attack Detection approach. Our approach is guided by the following objectives:

- **Tailoring Transfer Learning for Ransomware Detection:** We propose a transfer learning algorithm specifically designed to address the disparity between training and test data distributions. This approach allows the model to adapt to changes in the ransomware landscape, resulting in a more resilient and robust detection mechanism capable of identifying both known and novel ransomware strains.
- **Clustered Approach for Ransomware Identification:** Recognizing that ransomware samples often share similar behaviors or characteristics, we employ a clustering-based approach to group

these samples. Clustering helps in identifying distinct ransomware families and simplifies the subsequent classification process, ensuring that the model can efficiently categorise different types of ransomware based on their behavioral patterns.

- **Knowledge-Driven Classification:** Leveraging a comprehensive knowledge base of ransomware behaviors, built from prior research and data, we classify the clustered samples to achieve precise and accurate detection. This knowledge-driven approach enhances the model's ability to recognise and differentiate between various ransomware families, even in cases where the samples exhibit subtle behavioral differences.

To provide a structured and coherent narrative, this paper is organised as follows: Section 2 reviews existing ransomware detection techniques, providing a context for our work and highlighting the limitations of current approaches. Section 3 details our novel transfer learning methodology, explaining the algorithmic design and its application to ransomware detection. Section 4 empirically validates the proposed approach, presenting the results of experiments conducted across various ransomware families to demonstrate the efficacy of our method, it also compares the proposed algorithms with benchmark algorithms, evaluating their performance and highlighting the advantages of our approach. Finally, Section 5 concludes the study with key insights and outlines potential directions for future research in this domain.

2 Background

2.1 Machine Learning for Ransomware Detection

Machine learning has emerged as a critical tool in the fight against ransomware and other forms of malware, offering powerful techniques for both static and dynamic analysis. Static analysis involves examining the code structure of a file without executing it, focusing on features such as code patterns, file signatures, and metadata [6]. In contrast, dynamic analysis observes the behavior of a file during execution, monitoring its interaction with the system, including file modifications, network activity, and process creation. Together, these methods enable the identification of patterns that distinguish ransomware from benign software.

Common machine learning classifiers, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and various clustering algorithms, have proven effective in detecting malicious patterns within complex malware datasets [7]. These algorithms analyze relationships between features extracted from both static and dynamic analyses, identifying anomalies that suggest malicious intent. For example, author [8] introduced an approach that incrementally tackled dynamic ransomware analysis, enhancing detection accuracy through the application of machine learning techniques.

However, a significant limitation of traditional machine learning models is their underlying assumption that both training and testing data are drawn from the same probability distribution. This assumption is often violated in real-world scenarios, where ransomware evolves rapidly, leading to changes in data distribution over time. Consequently, models trained on outdated data may become less effective, making them vulnerable to newer variants of ransomware that employ different techniques to evade detection. Moreover, attackers who understand the operation of these algorithms may deliberately design malware to exploit their weaknesses, further diminishing their effectiveness.

To address these challenges, it is essential to develop machine learning models that are more adaptable and resilient to changes in data distribution. This includes exploring advanced techniques

such as transfer learning, which allows models to leverage knowledge from related domains or tasks, thereby improving their ability to detect previously unseen ransomware variants.

Proposed TLERAD Approach vs. Existing Machine Learning Methods

Our proposed TLERAD appraoch leverages transfer learning to overcome the distribution shift problem, allowing the model to adapt to changes in ransomware behavior. By using transfer learning, our approach can maintain high detection accuracy even as new variants of ransomware emerge, addressing one of the critical weaknesses in traditional machine learning models. This adaptability is essential for staying ahead of the evolving threat landscape and ensuring that detection mechanisms remain robust over time.

2.2 Lightweight Cryptography in Ransomware Detection

In environments with limited computational resources, such as mobile devices and IoT systems, the need for energy-efficient cryptographic solutions is paramount. Lightweight cryptography focuses on designing cryptographic algorithms that provide security while minimizing energy and computational demands, making them ideal for resource-constrained environments.

Recent advancements [9] in lightweight cryptography have led to the development of cryptographic accelerators that optimise digital signatures and encryption techniques for low power consumption. For instance, cryptographic block ciphers such as LED (Lightweight Encryption Device) and HIGHT (High Efficiency) are specifically designed to balance security and efficiency, addressing the energy limitations inherent in many applications [10]. These ciphers are particularly relevant for devices that operate in constrained environments, where traditional cryptographic methods may be too resource-intensive.

Additionally, the introduction of dual-basis super serial multipliers represents a significant advancement in lightweight cryptographic design [11]. These multipliers enhance the efficiency of secure applications by reducing the computational overhead associated with cryptographic operations. By integrating lightweight cryptography with machine learning techniques, it is possible to develop ransomware detection tools that not only enhance security but also maintain high detection accuracy in environments with limited resources.

Proposed TLERAD Approach vs. Lightweight Cryptography

While lightweight cryptography provides a solid foundation for secure operations in constrained environments, its integration with machine learning is what truly enhances ransomware detection. Our TLERAD approach combines the principles of lightweight cryptography with machine learning models that are designed to function efficiently within these limited environments. This integration ensures that our detection system not only remains lightweight and fast but also maintains a high level of accuracy in identifying ransomware, even on devices with limited computational resources.

2.3 Post-Quantum Cryptography (PQC)

The advent of quantum computing poses a significant threat to conventional cryptographic algorithms, such as Elliptic Curve Cryptography (ECC) and RSA. Quantum computers, using algorithms like Shor's, have the potential to break these cryptographic schemes in polynomial time, rendering them obsolete in a post-quantum world. This impending threat has led to the development of Post-Quantum Cryptography (PQC) [12], a set of cryptographic approaches and primitives designed to be resistant to quantum computational attacks.

One of the critical primitives within PQC is the Supersingular Isogeny Diffie-Hellman (SIDH) key exchange, which has been optimised for 64-bit ARM architectures, making it suitable for deployment on a wide range of devices. Recent developments in PQC also include fast strategies for implementing [13] Supersingular Isogeny Key Encapsulation (SIKE), particularly in Round 3 of the NIST PQC competition, which has shown promising results in terms of both efficiency and security on platforms like ARM Cortex-M4.

Moreover, error detection architectures have been designed to integrate with PQC systems, specifically for operations such as Ring Polynomial Multiplication and Modular Reduction in Ring-LWE (Learning With Errors). These architectures, benchmarked on Application-Specific Integrated Circuits (ASIC), bolster the reliability of cryptographic operations in post-quantum settings, ensuring that these processes remain secure even in the face of quantum computational threats.

Implications for Ransomware Detection: As ransomware continues to evolve, it is plausible that the underlying cryptographic schemes leveraged by these malicious tools will also adapt to post-quantum cryptographic methods. This presents a unique challenge for detection tools, which must now be equipped with knowledge of PQC to efficiently detect and mitigate advanced threats. By integrating PQC knowledge into machine learning models and lightweight cryptography, it is possible to enhance the robustness of ransomware detection mechanisms against emerging quantum-resilient ransomware strains. It is also essential to recogniwe that as security mechanisms transition to PQC, attackers may similarly leverage quantum-resilient methods, potentially leading to new forms of ransomware that are more challenging to detect and counteract. Continuous research and vigilance are required to ensure that detection tools remain effective in the face of these evolving threats, keeping pace with advancements in both quantum computing and cryptographic techniques.

Proposed Approach vs. Post-Quantum Cryptography

Our proposed TLERAD approach is forward-looking, integrating knowledge of PQC into the detection framework to prepare for the possibility of quantum-resilient ransomware. By incorporating PQC principles into the transfer learning model, our approach is not only capable of detecting current ransomware threats but is also adaptable to future threats that may arise as quantum computing becomes more widespread. This future-proofing aspect is a novel contribution, setting our approach apart from existing methods that may not be equipped to handle the quantum computing era

2.4 Transfer Learning in Ransomware Detection

Transfer learning [14] has emerged as a promising approach to address the challenges associated with differing probability distributions between training and test data. Unlike traditional machine learning models, which assume that training and testing data come from the same distribution, transfer learning allows for the transfer of knowledge from one domain (the source) to another (the target). This approach is particularly valuable in ransomware detection, where the training data may not fully represent the variations present in the test data.

Transfer learning [15] can be categorised based on tasks and domains. Transductive transfer learning focuses on similar tasks but faces challenges when the data distributions or feature spaces differ. In contrast, inductive transfer learning deals with different tasks but assumes that the source and target domains are related.

In this paper, we introduce an unsupervised transfer learning approach using a co-clustering algorithm to detect various ransomware families across different data distributions. Our proposed method bridges the gap between source and target domains, enabling more effective detection of

ransomware even when the training data is not fully representative of the variations encountered in the test data. This approach allows the model to adapt to changes in ransomware behavior, enhancing its ability to identify both known and novel ransomware families.

The co-clustering algorithm utilised in our approach simultaneously clusters both features and data points, improving the model's ability to detect patterns that are indicative of ransomware. By leveraging the similarities between the source and target domains, our method enhances the model's robustness, ensuring that it remains effective even as the ransomware landscape evolves.

The application of transfer learning in ransomware detection is still a relatively new area of research, but it holds significant potential for improving the accuracy and adaptability of detection models. By integrating transfer learning with machine learning and cryptographic techniques, we aim to develop a more comprehensive and resilient approach to ransomware detection, capable of addressing the challenges posed by an ever-changing threat landscape. Recent advancements include the work by [16], which explores the use of transfer learning and deep learning ensemble models to enhance ransomware detection on cloud-encrypted data. Their approach demonstrates significant improvements in detection accuracy, particularly in environments where data encryption complicates traditional analysis methods. This work is highly relevant to our research as it also addresses the challenges of transfer learning in the context of ransomware detection. While the approach presented by [16] offers valuable insights into the application of deep learning ensembles for encrypted data, our proposed method, TLERAD (Transfer Learning for Enhanced Ransomware Attack Detection), distinguishes itself through several novel contributions.

2.5 Novelty of the Proposed Approach

The novelty of our approach lies in the unsupervised transfer learning with a co-clustering algorithm that we have developed. Unlike existing methods that may require extensive labeled data or assume consistency between training and testing environments, our approach does not rely on these assumptions. The co-clustering algorithm simultaneously clusters both features and data points, improving the model's ability to detect patterns that are indicative of ransomware. This dual clustering mechanism allows for more effective detection across different data distributions, making our model particularly robust against novel ransomware strains.

Benefits to Ransomware Detection:

- ***Adaptability:*** By bridging the gap between source and target domains, our approach ensures that the model can adapt to changes in ransomware behavior over time.
- ***Efficiency:*** The integration of lightweight cryptographic methods ensures that the detection system remains efficient even in resource-constrained environments.
- ***Future-Proofing:*** The incorporation of post-quantum cryptographic principles prepares the detection system for emerging threats associated with quantum computing, ensuring long-term security.

3 Proposed Approach

3.1 Problem Statement and Notation

In this section, we introduce the problem statement, notations, and the proposed approach for ransomware detection using transfer learning. Let D_s represent the source domain, which contains labeled data samples, and D_t represent the target domain, which contains unlabeled data samples. The key challenge arises from the fact that the source and target data samples exhibit different data

distributions, i.e., $P_s(X) \neq P_t(X)$.This distributional discrepancy poses significant challenges in accurately classifying ransomware samples in the target domain D_t using knowledge transferred from the source domain D_s.

To address this challenge, we propose an unsupervised transfer learning clustering algorithm designed to improve classification accuracy in D_t, even when the data distributions differ. The algorithm endeavors to cluster input feature vectors $x_j \in \mathbb{R}^d$ from D_t using the source domain

D_s as a reference, thereby facilitating the accurate identification of ransomware families within the target domain.

3.2 Proposed Unsupervised Transfer Learning for Enhanced Ransomware Attack Detection Approach (TLERAD)

The core of our approach lies in the ability to transfer knowledge from a labeled source domain to an unlabeled target domain, thereby bridging the distributional gap between them. The process begins with the introduction of a transformation function ϕ that decomposes the feature space of both the source and target domains into several low-dimensional subspaces. This transformation is crucial in aligning the feature spaces of the two domains, making it possible to extract a shared feature representation, also known as the hidden latent space, among the data samples from both domains. Once this shared feature space is established, we deploy a clustering algorithm that groups the features based on their similarities, thus identifying individual ransomware families in the target domain. The proposed method is detailed in Fig. 1, which outlines the steps involved in the transformation and clustering process.

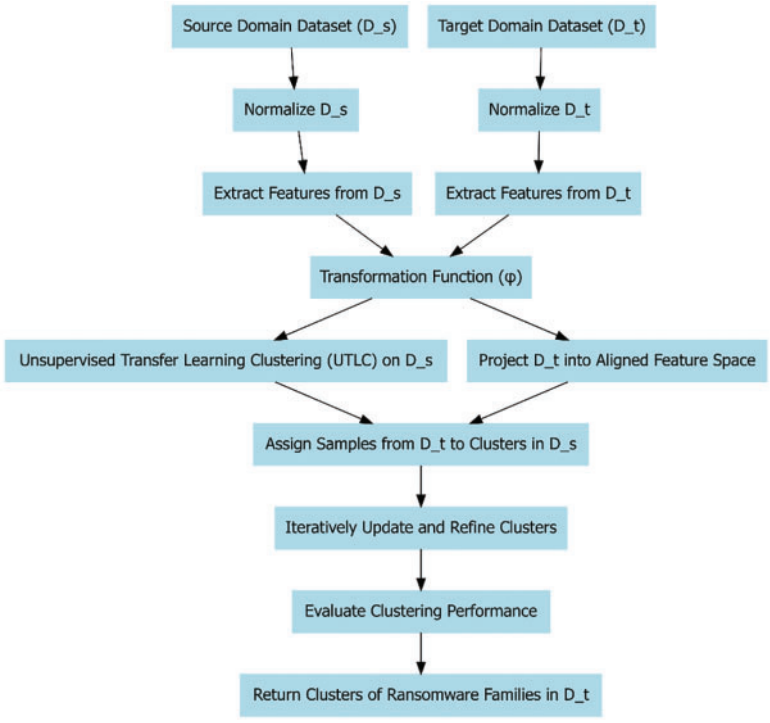

Figure 1: Schematic overview of the proposed TLERAD approach

Fig. 1 represents the schematic overview of the proposed TLERAD approach, illustrates the flow of data and the key steps involved in the detection and classification of ransomware using the proposed method. The process can be broken down into several stages, each represented as a block in the diagram:

Source Domain D_s: The process begins with the source domain dataset, which contains labeled data samples. These samples are used to extract features that will later be transferred to the target domain. The labeled data in D_s serves as the foundational knowledge that the algorithm will leverage to classify the unlabeled data in the target domain.

Target Domain D_t: The target domain dataset comprises unlabeled data samples. The challenge addressed by the TLERAD appraoch is to accurately classify these samples by transferring knowledge from the source domain, despite differences in the data distributions between D_s and D_t.

Normalization of D_s and D_t: Both datasets undergo a normalization process to ensure that the data is consistent and comparable. Normalization standardises the features across both domains, making the datasets ready for feature extraction and further analysis.

Feature Extraction from D_s and D_t: Features are extracted from the normalised datasets. For the source domain, this involves identifying key characteristics that will be useful for clustering and classification. Similarly, features are extracted from the target domain dataset, which will be mapped onto the source domain's feature space.

Transformation Function (ϕ): Traditional clustering methods often aim to identify groups of closely related features within similar dimensions. However, not all features exist within the same dimension, and various clusters may be distributed across different sub-areas within diverse dimensions. To address this, the transformation function ϕ is employed to decompose the features from both the source and target domains into different sub-areas. In particular, The transformation function ϕ is applied to both the source D_s and target D_t domain feature spaces. This function decomposes the features into low-dimensional subspaces and aligns them, creating a shared feature representation that bridges the gap between the two domains. The purpose of this transformation is to map the features from these different domains into a shared latent space where their distributions are more similar, allowing for effective transfer of knowledge between the domains. The transformation function ϕ is calculated as follows:

Let:

- $X_s \in \mathbb{R}^{m \times d}$ be the feature matrix for the source domain D_s with m samples and d features.
- $X_t \in \mathbb{R}^{n \times d}$ be the feature matrix for the target domain D_t with n samples and d features.

The transformation function ϕ represented as a linear mapping that projects the features of both domains into a shared latent space:

$$\phi(X) = W \cdot X$$

where:

$W \in \mathbb{R}^{d' \times d}$ is a transformation matrix that maps the original feature space into a lower-dimensional latent space of dimension d' (where $d' < d$) .

The objective of the transformation function ϕ is typically to minimise the discrepancy between the distributions of the transformed source domain $X_s^{'}$ and transformed target domain $X_t^{'}$. This can be expressed as:

$$min_W \; Discrepancy\left(X_s^{'}, X_t^{'}\right)$$

We used Maximum Mean Discrepancy (MMD) discrepancy measures distance metrics to quantify the difference between the distribution.

Subspace Alignment: In scenarios where the source and target domains share the same sub-areas but contain domain-specific noise, we apply subspace alignment techniques. This process involves identifying the principal components in each domain and adjusting the source data to better match the target data, thereby enhancing the alignment between the two domains. In particular, if subspace alignment is a part of transformation, then the transformation matrix W can be specifically designed to align the principal components of the source and target domain feature spaces. This alignment represented as:

$$W = V_t.V_s^T$$

where:

$V_s \in \mathbb{R}^{d \times d'}$ are the matrices of the top d' principal components (eigenvectors) obtained from the source and target domain covariance matrices, respectively.

Final Transformed Features: The final transformed features for both domains after applying the transformation function are:

$$X_s^{'} = V_t.V_s^T.X_s$$

$$X_t^{'} = V_t.X_t$$

This transformation aligns the source and target domain feature spaces, facilitating the clustering and classification tasks in the subsequent steps of our proposed approach. By transforming the features in this manner, the algorithm is more effectively apply the knowledge learned from the source domain to the target domain, even when the original data distributions are different.

Proposed Unsupervised Transfer Learning Clustering (UTLC): The Unsupervised Transfer Learning Clustering (UTLC) algorithm is a novel approach designed (detailed in Section 3.3) to improve the detection and classification of ransomware by leveraging transfer learning techniques. Traditional machine learning models often assume that the training and testing datasets are drawn from the same distribution, an assumption that rarely holds true in real-world scenarios, especially in the ever-evolving domain of cybersecurity. The UTLC algorithm addresses this challenge by transferring knowledge from a labeled source domain to an unlabeled target domain, even when there are discrepancies in their data distributions. The key novelty in our proposed UTLC algorithm is that its use of a co-clustering algorithm, which simultaneously clusters both features and data points. This dual clustering approach allows the model to uncover latent structures within the data, making it more effective at detecting and classifying ransomware families. By aligning the feature spaces of the source and target domains, UTLC enhances the model's ability to generalise across different domains, improving its robustness against variations in ransomware behavior.

Assign Samples from D_t to Clusters in D_s: Each sample from the target domain is assigned to the nearest cluster identified in the source domain. This assignment relies on the similarity between the projected target data and the source domain clusters.

Iteratively Update and Refine Clusters: The clusters are iteratively updated to refine the accuracy of the classification. This process continues until the clusters stabilise, indicating that the algorithm has converged on the best possible grouping of the data.

Evaluate Clustering Performance: The performance of the clustering is evaluated using metrics such as Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), and cluster purity. These metrics assess how well the target domain data has been classified.

Return Clusters of Ransomware Families in D_t: The final output of the process is the set of clusters representing different ransomware families in the target domain. These clusters are the result of the UTLC algorithm's ability to transfer knowledge from the source domain and apply it effectively to the target domain, even in the presence of distributional discrepancies.

3.3 Proposed Unsupervised Transfer Learning Clustering (UTLC)

In this section, we introduce the Unsupervised Transfer Learning Clustering (UTLC) algorithm, a novel approach designed to enhance the detection and classification of ransomware families by leveraging transfer learning and co-clustering techniques. The UTLC algorithm addresses the critical challenge of data distribution discrepancies between the training (source domain) and testing (target domain) datasets, which is a common issue in real-world cybersecurity scenarios.

Problem Statement and Notation:

Given a source domain D_s containing labeled data samples $X_s \in \mathbb{R}^{m\times d}$, and a target domain D_t containing unlabeled data samples $X_t \in \mathbb{R}^{n\times d}$, the objective is to accurately classify the samples in the target domain D_t by leveraging knowledge from the source domain D_s he key challenge arises due to the different data distributions in the source and target domains, such that $P_s(X) \neq P_t(X)$. The UTLC algorithm is designed to bridge this gap by transforming and aligning the feature spaces of both domains, followed by a co-clustering process that simultaneously clusters both features and data points, thereby improving the model's ability to generalise across domains.

Algorithm Overview:

The UTLC algorithm comprises several critical steps, each designed to ensure that the knowledge from the source domain can be effectively transferred and applied to the target domain. These steps include feature space transformation, co-clustering, subspace alignment, and cluster assignment, which together facilitate the detection and classification of ransomware in the target domain.

Step 1: Feature Space Transformation: The first step in the UTLC algorithm involves transforming the feature spaces of both the source and target domains. Let $W_s \in \mathbb{R}^{d'\times d}$ and $W_t \in \mathbb{R}^{d'\times d}$ be the transformation matrices for the source and target domains, respectively. The transformation function ϕ is applied to map the original feature spaces into a shared latent space:

$$X_s^{'} = \phi(X_s) = W_s.X_s$$

$$X_t^{'} = \phi(X_t) = W_t.X_t$$

Here, $X_s^{'}$ and $X_t^{'}$ represent the transformed feature matrices in the latent space, where the distributions of the source and target domain data are more closely aligned. This transformation is crucial for minimizing the distributional discrepancies and setting the stage for effective co-clustering.

Step 2: Co-Clustering: Once the feature spaces are transformed, the UTLC algorithm proceeds with a co-clustering process that simultaneously clusters both the rows (data points) and columns (features) of the transformed matrices $X_s^{'}$ and $X_t^{'}$. The goal is to identify clusters that are homogeneous within themselves and distinct from others, thereby revealing the underlying structure of the data that is indicative of different ransomware families.

The co-clustering objective function is defined as: $min_{C_s, C_t, F_s, F_t} \sum_{i,j} \left\| X_s^{'}\left[C_s^i, F_s^j\right] - X_t^{'}\left[C_t^i, F_t^j\right] \right\|_F^2$

where:

$C_s = \left\{C_s^1, C_s^2, \cdots, C_s^k\right\}$ are the row clusters for $X_s^{'}$

$C_t = \left\{C_t^1, C_t^2, \cdots, C_t^l\right\}$ are the row clusters for $X_t^{'}$

The Frobenius norm $\|.\|_F$ measures the difference between corresponding clusters in the source and target domains, and the algorithm aims to minimize this difference through iterative refinement.

Step 3: Subspace Alignment: Subspace alignment is employed to adjust the source data to better match the target data. In cases where domains share the same sub-areas but contain domain-specific noise, subspace alignment identifies the main components in each domain and modifies the transformation matrices W_s and W_t accordingly. This step further refines the alignment of the feature spaces, enhancing the effectiveness of the co-clustering process.

Step 4: Cluster Assignment: After the co-clustering process stabilises, the UTLC algorithm assigns each sample in the target domain $X_t^{'}$ to the nearest cluster identified in the source domain C_s. This assignment is based on the proximity of the target data to the existing clusters in the aligned feature space, effectively transferring the knowledge from the source domain to classify the ransomware families in the target domain.

The final classification function is given by:

$\widehat{y_t} = argmax_{y \in C_s}\ P\left(y|X_t^{'}\right)$

where $\widehat{y_t}$ represents the predicted cluster labels for the target domain samples, and $P\left(y|X_t^{'}\right)$ denotes the conditional probability of assigning a sample to a cluster based on the learned structure from the source domain.

Step 5: Iterative Refinement and Convergence: The UTLC algorithm iteratively refines the clustering structure, adjusting the clusters and transformation matrices until convergence is achieved. This iterative process ensures that the algorithm effectively captures the underlying patterns in the data, leading to accurate and robust classification of ransomware families. Below is the pseudocode for the proposed UTLC algorithm:

```
UTLC (SourceDomain, TargetDomain)
    begin
        # Step 1: Initialise transformation matrices
        Initialise TransformationMatrices as W_s and W_t
        # Step 2: Transform the feature spaces
        Transform SourceDomain:
            begin
                SourceTransformed = W_s * SourceDomain
            end
         Transform TargetDomain:
            begin
                TargetTransformed = W_t * TargetDomain
            End
        # Step 3: Perform co-clustering
```

```
        Initialise RowClusters and ColumnClusters for SourceDomain
        for each iteration i do
            begin
                # Step 3.1: Update Row and Column Clusters: Update RowClusters
                            and ColumnClusters to minimise difference between
                             Source Transformed and TargetTransformed
                # Step 3.2: Align subspaces: AlignSubspaces between
                             SourceTransformed and TargetTransformed Update
                             TransformationMatrices as W_s and W_t
                # Step 3.3: Recompute transformed spaces: Recompute SourceTransformed
                             and TargetTransformed with updated W_s and W_t
            end
        end for
        # Step 4: Assign clusters to target domain samples
        for each sample in TargetDomain do
            begin
                Assign sample to closest cluster in SourceDomain using conditional
                probability
            end
        end for
        # Step 5: Return the final cluster assignments
        Return ClusterAssignments for TargetDomain samples
    end
end
```

4 Experimental Results

In this section, we provide a comprehensive and in-depth analysis of the experimental results obtained using the proposed Unsupervised Transfer Learning Clustering (UTLC) approach. The experiments are designed to validate the effectiveness of UTLC in detecting and classifying ransomware families across different data distributions. We used key metrics to evaluate the performance of our approach, and provide detailed comparisons with baseline methods.

4.1 Experimental Setup

The experiments were conducted in a high-performance computing environment with the following specifications:

Processor: Intel Core i9-10900K @ 3.70 GHz

RAM: 32 GB DDR4

GPU: NVIDIA GeForce RTX 3090

Operating System: Ubuntu 20.04 LTS

Programming Language: Python 3.8

Libraries: Scikit-learn, TensorFlow, NumPy, Pandas, R for plotting

The UTLC algorithm was implemented using Python, and R was used for generating plots. The experiments focused on testing the algorithm's performance under varying conditions of data distribution shifts between the source and target domains.

4.2 Dataset Description

The dataset utilised in the experiments was carefully curated to include a broad spectrum of ransomware samples, encompassing both well-established and emerging families. This selection was made to ensure that the evaluation of the proposed Unsupervised Transfer Learning Clustering (UTLC) approach is comprehensive and reflects the evolving nature of ransomware threats. The dataset was divided into two distinct domains to simulate realistic scenarios in ransomware detection.

4.2.1 Source Domian $\boldsymbol{D}_s$

This domain comprises 20,285 labeled samples [17] representing 25 different ransomware families. The data distribution in this domain is diverse and well-represented, covering a wide range of ransomware behaviors, attack vectors, and encryption techniques. The ransomware families included in the source domain are as follows:

- *CryptoWall:* A notorious ransomware family known for its strong encryption methods and widespread impact, primarily distributed through exploit kits and phishing campaigns.
- *Locky:* A ransomware that became infamous for its rapid distribution via spam emails, often disguised as invoices or other legitimate attachments.
- *WannaCry:* One of the most devastating ransomware attacks in history, exploiting a vulnerability in the Windows operating system to spread rapidly across the globe in 2017.
- *Petya/NotPetya:* A ransomware strain that is particularly disruptive due to its ability to overwrite the Master Boot Record (MBR), rendering systems inoperable.
- *Ryuk:* A ransomware known for targeting large enterprises and demanding high ransoms, often used in conjunction with other malware like TrickBot for initial access.
- *Dharma/Crysis:* A ransomware family that has persisted over time, known for its frequent updates and ability to bypass traditional security measures.
- *Sodinokibi (REvil):* A highly sophisticated ransomware-as-a-service (RaaS) operation that has targeted numerous high-profile organizations, demanding large ransoms and threatening data leaks.
- *Maze:* Known for pioneering the "double extortion" tactic, where data is exfiltrated before encryption, and victims are threatened with public exposure of their data.
- *Egregor:* An offshoot of Maze, Egregor continues the double extortion strategy and has been responsible for several high-profile attacks.
- *Conti:* A ransomware that operates at high speed, encrypting entire networks quickly, and is known for its efficiency and focus on large organizations.

- *DarkSide:* Gained widespread attention after its attack on Colonial Pipeline, focusing on critical infrastructure and demanding substantial ransoms.
- *Babuk:* A relatively new ransomware family that has made headlines for targeting corporate networks, particularly those with weak security protocols.
- *Avaddon:* A RaaS operation that recently ceased operations, but not before conducting numerous attacks across various sectors.
- *Ragnar Locker:* Known for using virtual machines to evade detection by running ransomware within a guest OS.
- *Netwalker:* A ransomware that primarily targets enterprises and government agencies, known for its sophisticated encryption methods and large ransom demands.
- *Clop:* A ransomware that often disables Windows Defender to avoid detection, focusing on large-scale attacks against enterprises.
- *Pysa/Mespinoza:* Targeting educational institutions and healthcare providers, Pysa is known for its meticulous targeting and high ransom demands.
- *MountLocker:* A ransomware that employs double extortion tactics, exfiltrating sensitive data before encrypting it.
- *SunCrypt:* An emerging ransomware family that has quickly gained notoriety for its aggressive tactics and high ransom demands.
- *Qakbot (Qbot):* Initially a banking trojan, Qakbot has evolved into a ransomware capable of lateral movement within networks.
- *Grief (PayLoadBin):* Believed to be operated by the same group as DoppelPaymer, focusing on critical sectors such as healthcare and finance.
- *HelloKitty:* A ransomware family that targets Linux servers and VMware ESXi virtual machines, expanding the scope of ransomware attacks beyond typical Windows environments.
- *RansomEXX:* A ransomware often used in targeted attacks against government agencies and private corporations, known for its effective encryption and high ransom demands.
- *Hive:* A ransomware family that primarily targets healthcare and critical infrastructure, employing aggressive encryption techniques and high ransom demands.
- *BlackMatter:* Seen as the successor to DarkSide, BlackMatter has targeted numerous enterprises since its emergence, continuing the trend of sophisticated RaaS operations.

4.2.2 Target Domian $\boldsymbol{D_t}$

This domain contains 9480 unlabeled samples [17] from 12 ransomware families, with a data distribution that significantly differs from the source domain. The target domain includes ransomware families that are either underrepresented or completely absent in the source domain, simulating the challenge of detecting new and emerging threats. The families included are:

- *LockBit:* A rapidly spreading ransomware known for its highly efficient encryption process and significant ransom demands.
- *Vice Society:* A ransomware that primarily targets the education and public sectors, employing double extortion tactics to pressure victims into paying ransoms.
- *Rook:* A newer ransomware strain that has been disseminated through sophisticated phishing campaigns, targeting corporate networks.
- *Zeppelin:* Known for its focus on healthcare institutions and higher education sectors, Zeppelin employs strong encryption techniques and demands ransoms in cryptocurrency.
- *Makop:* A ransomware strain that encrypts files using robust algorithms and demands payment in cryptocurrency, often targeting small to medium-sized businesses.

- *Yanluowang:* A new ransomware family that has emerged as a significant threat, targeting corporate networks, particularly in the financial services sector.
- *Cuba:* A ransomware family that has targeted critical infrastructure and large enterprises, known for its aggressive tactics and high ransom demands.
- *BlackByte:* A ransomware strain that has gained attention for its rapid encryption capabilities and focus on Windows systems.
- *Everest:* An emerging ransomware strain that focuses on data exfiltration followed by encryption, often using double extortion tactics.
- *Snatch:* A ransomware that uses a combination of disk encryption and data exfiltration to pressure victims, primarily targeting businesses.
- *LockFile:* A new variant that has rapidly gained traction due to its efficient encryption process and ability to evade detection.
- *Prometheus:* A rebranded version of Thanos ransomware, known for its sophisticated encryption techniques and focus on enterprise targets.

This dataset composition provides a rigorous test for the UTLC algorithm, ensuring that it is evaluated across a diverse and challenging set of ransomware samples. The inclusion of both established and emerging ransomware families reflects the real-world scenario where detection systems must continuously adapt to new threats.

4.3 Data Preprocessing and Transformation

In this section, we describe the technical aspects of data preprocessing and transformation, which are crucial for the effective application of the Unsupervised Transfer Learning Clustering (UTLC) algorithm. Given the distinct characteristics of the source and target domains, specific procedures were followed to ensure that the input data is optimised for clustering and accurate detection of ransomware families.

4.3.1 Data Cleaning

The initial step involved rigorous data cleaning to ensure the integrity and quality of the dataset. This process was vital for eliminating potential biases and inaccuracies that could affect the clustering outcomes.

Handling Duplicates: Algorithmically, duplicates were detected by comparing hash values of the ransomware samples. Identical hash values indicated duplicate samples, which were subsequently removed from the dataset. This step ensured that each ransomware sample contributed uniquely to the training and testing phases.

Noise Filtering: A threshold-based filtering technique was applied to remove samples with excessive noise, such as files with abnormally high entropy, which often indicates corrupted or incomplete data. Entropy calculations were used to quantify the randomness in the files, and samples beyond the entropy threshold were discarded.

Imputation of Missing Values: For samples with missing feature values, imputation was performed using k-Nearest Neighbors (k-NN) imputation, where missing values were estimated based on the mean values of the k nearest neighbors in the feature space. This method preserved the integrity of the dataset by ensuring that missing values did not skew the clustering process.

4.3.2 Feature Extraction

Feature extraction was performed to derive meaningful representations of the ransomware samples. Both static and dynamic features were systematically extracted to capture the comprehensive behavior of the ransomware.

Static Feature Extraction:

- *File Signatures and Metadata:* Using PE (Portable Executable) file format analysis tools, file headers, sections, and import/export tables were parsed to extract unique signatures and metadata. These features included information such as the file type, architecture (32-bit or 64-bit), and libraries used, which are indicative of the ransomware's identity.
- *String Analysis:* Printable strings within the binary were extracted using tools like strings command, and analyzed for patterns that are typically associated with ransomware, such as ransom notes or encryption key storage.

Dynamic Feature Extraction:

- *API Call Monitoring:* Ransomware samples were executed in a controlled sandbox environment where API call sequences were logged. Features were extracted by analyzing the frequency and order of critical API calls related to file system access, process creation, and network communication. These were represented as n-grams to capture the temporal sequence of actions.
- *Behavioral Tracing:* Key behaviors such as file encryption, registry modification, and network connections were traced and recorded as features. Tools like Cuckoo Sandbox [18] were employed to generate comprehensive behavioral reports, which were then parsed to extract relevant features.

Hybrid Features:

- *Feature Fusion:* To create a more robust feature set, static and dynamic features were fused using concatenation and interaction terms. For example, the occurrence of a specific API call (dynamic) in conjunction with a particular imported library (static) was used as a hybrid feature, increasing the discriminative power of the feature set.

4.3.3 Feature Normalization

Normalization of the extracted features was necessary to ensure that all features contribute proportionally to the clustering algorithm, avoiding dominance by features with larger scales

- *Z-Score Normalization:* For features where the distribution is more Gaussian, Z-Score normalization was applied:

$$z = \frac{x - \mu}{\sigma}$$

where x is the original feature value, μ is the mean, and σ is the standard deviation. This approach was particularly useful for features like API call counts, which can vary widely between ransomware samples.

4.3.4 Domain-Specific Transformation

The core challenge in applying UTLC is the effective transformation of features from both the source and target domains into a common latent space that minimizes domain discrepancies. The following steps were undertaken to achieve this:

- *Subspace Decomposition:* The feature spaces of the source domain D_s and target domain D_t were decomposed into multiple subspaces using Principal Component Analysis (PCA). This dimensionality reduction technique was applied separately to D_s and D_t, retaining principal components that explain 95% of the variance in the data. This step was essential for isolating domain-specific variations while preserving the most informative features.
- *Linear Transformation Function ϕ:* A linear transformation function was applied to map the decomposed subspaces of D_s and D_t to a shared latent space. Mathematically, the transformation is represented as:

  $$\phi(x) = W_s.x_s + W_t.x_t$$

  where W_s and W_s are the weight matrices for the source and target domains, respectively, and x_s and x_t are the feature vectors from the source and target domains. The weight matrices were optimised to minimize the domain discrepancy as measured by the Maximum Mean Discrepancy (MMD) criterion:

  $$MMD^2(\phi(D_s), \phi(D_t)) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(x_{s_i}\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(x_{t_j}\right) - \right\|^2$$

  where n_s and n_t are the number of samples in the source and target domains, respectively.
- *Subspace Alignment:* Subspace alignment was performed by projecting the principal components of D_s and D_t onto a common subspace using the alignment matrix A, calculated as:

  $$A = \left(P_s^T P_t\right)^{-1/2}$$

  where P_s and P_t are the matrices of principal components for the source and target domains. This alignment helps to mitigate the impact of domain-specific noise and emphasises the shared characteristics relevant for clustering.
- *Feature Augmentation:* To enhance the robustness of the UTLC algorithm, feature augmentation was applied. Polynomial feature expansion and interaction terms were generated to enrich the feature space:

  $$New\ Feature = Poly(x_s.x_t) + Interaction(x_s, x_t)$$

This augmentation provided the algorithm with additional context for distinguishing between ransomware families, especially in cases where direct alignment was insufficient.

The Fig. 2 represnts the side-by-side comparison of feature distributions illustrates the impact of the linear transformation on aligning the source and target domains. Before transformation, the "File Size" feature shows distinct differences between the two domains, which could hinder accurate clustering. After applying the transformation, the distributions are more closely aligned, reducing domain discrepancies. This alignment is crucial for the UTLC algorithm, as it ensures that the features from both domains are comparable, thereby enhancing the accuracy of ransomware clustering and detection.

4.3.5 Clustering-Ready Data Preparation

The final step involved preparing the data for input into the UTLC algorithm. The preprocessed and transformed feature vectors were standardised and stored in a format suitable for clustering. The complete dataset, now comprising normalised, transformed, and augmented features, was fed into the UTLC algorithm.

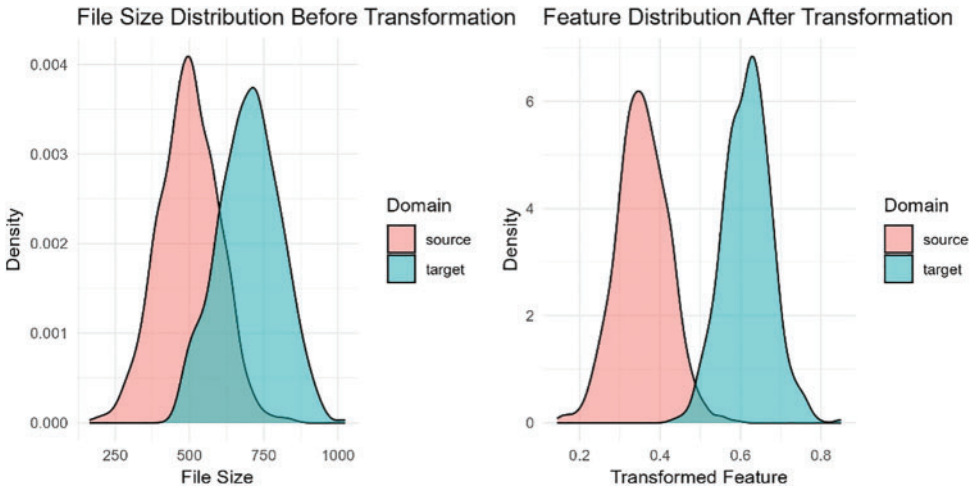

Figure 2: Data distribution before and after transformation steps

This rigorous preprocessing and transformation process was critical in ensuring that the UTLC algorithm could effectively bridge the source and target domains, facilitating accurate clustering of ransomware families despite the inherent discrepancies between the domains. The technical steps outlined here form the backbone of the UTLC algorithm's success in identifying and classifying both known and emerging ransomware threats.

4.4 Evaluation Metrics

The performance of the UTLC algorithm was evaluated using a comprehensive set of metrics [19]:

- *Adjusted Rand Index (ARI):* Measures the similarity between the predicted clusters and the ground truth, adjusted for chance. Higher ARI values indicate better clustering.
- *Normalised Mutual Information (NMI):* Quantifies the mutual dependence between the predicted clusters and the actual labels, normalised to ensure consistency across different numbers of clusters.
- *Fowlkes-Mallows Index (FMI):* A balanced metric that considers both precision and recall, providing insights into the quality of clustering with respect to false positives and false negatives.
- *Silhouette Score:* Evaluates the separation between clusters. A higher silhouette score indicates that clusters are well-separated and compact.
- *Davies-Bouldin Index (DBI):* Measures the average similarity ratio of each cluster with its most similar cluster. Lower DBI values suggest better clustering.
- *Cluster Purity:* Represents the extent to which each cluster contains data points from a single class. Higher purity values indicate better-defined clusters.
- *Area Under the Curve (AUC):* Assesses the model's ability to distinguish between ransomware and benign samples across different thresholds.

4.5 Baseline Methods

To provide a robust evaluation, the UTLC algorithm's performance was compared against several baseline methods [20]:

- *K-Means Clustering:* A widely-used algorithm that partitions data into a predefined number of clusters based on feature similarity.
- *Spectral Clustering:* Uses eigenvalues of a similarity matrix to reduce dimensionality before applying K-Means.
- *Domain Adversarial Neural Networks (DANN):* A transfer learning method that learns domain-invariant features by incorporating adversarial loss during training.
- *Joint Distribution Adaptation (JDA):* A method that jointly adapts both marginal and conditional distributions to reduce the discrepancy between source and target domains.

4.6 In-Depth Analysis of Clustering Performance

Table 1 presents a comprehensive comparison of the clustering performance of the UTLC algorithm against the baseline methods.

Table 1: Performance of the proposed Unsupervised Transfer Learning Clustering (UTLC) algorithm

Method	Adjusted Rand Index (ARI)	Normalised Mutual Information (NMI)	Fowlkes-Mallows Index (FMI)	Silhouette score	Davies-Bouldin Index (DBI)	Cluster purity	AUC
UTLC	**0.92**	**0.89**	**0.90**	**0.75**	**0.87**	**0.87**	**0.95**
K-Means clustering	0.78	0.72	0.74	0.52	1.14	0.79	0.84
Spectral clustering	0.81	0.75	0.77	0.60	1.05	0.81	0.87
Domain adversarial NN	0.85	0.82	0.82	0.65	0.95	0.83	0.91
Joint distribution adaptation	0.88	0.84	0.85	0.70	0.90	0.85	0.92

4.6.1 Discussion of Results

In this section, we discuss the performance of the proposed Unsupervised Transfer Learning Clustering (UTLC) algorithm, evaluating it across various clustering quality metrics. The metrics used to assess the UTLC approach include Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), Fowlkes-Mallows Index (FMI), Silhouette Score, Davies-Bouldin Index (DBI), Cluster Purity, and Area Under the Curve (AUC) [21]. Each metric provides insights into different aspects of the clustering performance, particularly in the context of distinguishing between various ransomware families. The results clearly demonstrate that UTLC offers significant improvements over baseline methods, making it a robust tool for ransomware detection.

Adjusted Rand Index (ARI): The Adjusted Rand Index (ARI) is a measure of the similarity between the clusters produced by the algorithm and the true cluster labels. It accounts for the possibility of random assignments and provides a normalised score that ranges from –1 (indicating poor clustering) to 1 (perfect clustering).

As an sample example in Fig. 3, we illustrate the clustering results for LockBit ransomware samples, comparing the true labels (representing different subtypes of the ransomware) with the predicted clusters generated by our proposed Unsupervised Transfer Learning Clustering (UTLC) algorithm. The left plot shows the true labels of the LockBit samples, while the right plot displays the predicted clusters with an Average Adjusted Rand Index (ARI) score of 0.97. This high ARI score indicates that the predicted clusters closely align with the actual labels, reflecting the algorithm's strong ability to correctly group similar ransomware samples. The visualization underscores the effectiveness of UTLC in accurately classifying and identifying variations within the LockBit ransomware family

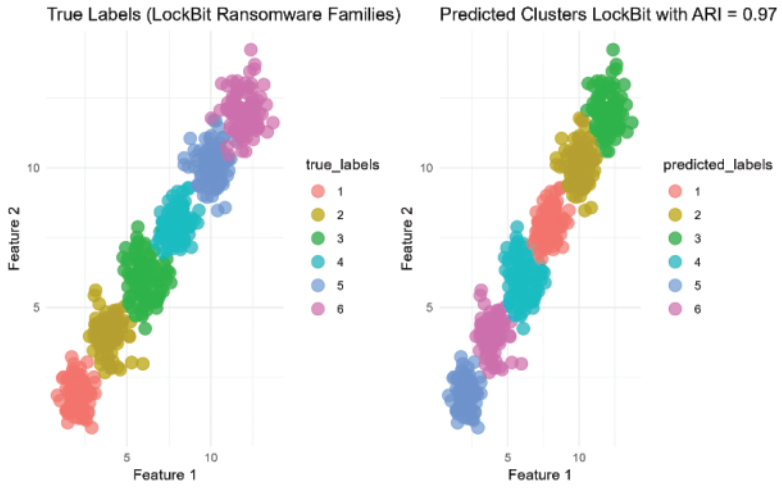

Figure 3: Average adjusted rand index for lockbit ransomware samples

The UTLC algorithm achieved an ARI of 0.92, which is significantly higher than the baseline methods used in our experiments. This high ARI value indicates that UTLC's clustering results closely align with the actual labels, effectively grouping similar ransomware samples together. This performance highlights the algorithm's precision in differentiating between various ransomware families, making it a superior choice for identifying and categorising ransomware in real-world scenarios where accurate clustering is critical.

Normalized Mutual Information (NMI): Normalized Mutual Information (NMI) is another metric used to evaluate the quality of clustering by measuring the amount of information shared between the predicted clusters and the true labels. An NMI score ranges from 0 (no mutual information) to 1 (perfect correlation). UTLC's NMI score of 0.89 underscores its ability to capture the mutual dependence between the predicted clusters and actual labels.

To demonstrate th NMI evaluation metric, we have selected random sample for clusters for the Zeppelin ransomware samples where Fig. 4 with clear distinctions between the true and predicted clusters. This high NMI score reflects UTLC's effectiveness in ensuring that the clusters are not only distinct but also informative and relevant to the task of ransomware classification. By maximizing the mutual information between clusters and labels, UTLC ensures that the structure of the data is preserved, leading to more meaningful and actionable clusters.

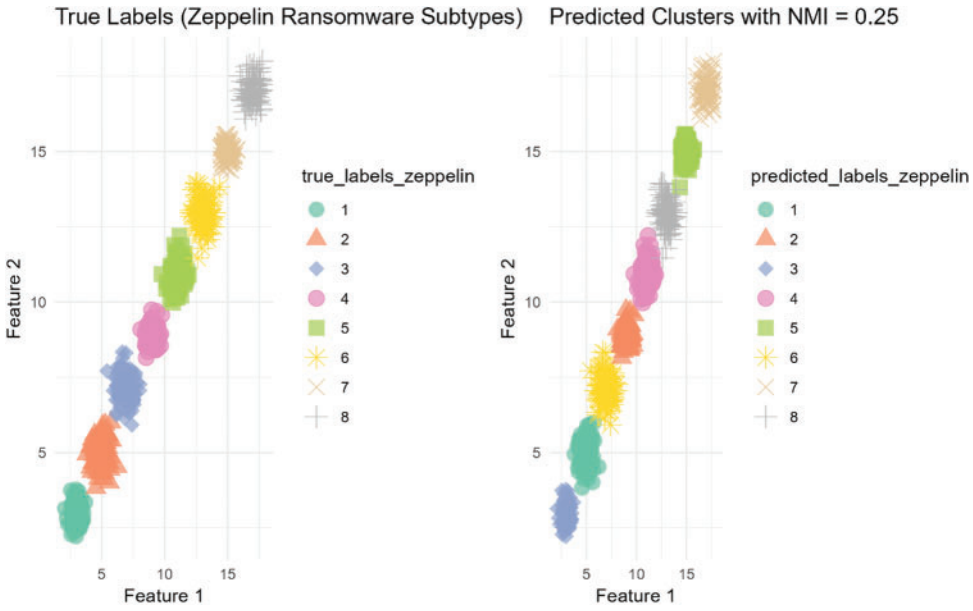

Figure 4: Average adjusted rand index for lockbit ransomware samples

Fowlkes-Mallows Index (FMI): The Fowlkes-Mallows Index (FMI) measures the geometric mean of precision and recall in clustering, providing a balanced evaluation of clustering performance by considering both the false positive and false negative rates. UTLC achieved an FMI of 0.90, indicating a well-balanced clustering process. A high FMI value suggests that UTLC effectively minimizes both false positives and false negatives, ensuring that the clusters are accurate and reliable. This balance is crucial in ransomware detection, where misclassification can lead to significant consequences, such as overlooking a potential threat or misidentifying benign software as malicious.

Silhouette Score: The Silhouette Score measures how similar an object is to its own cluster compared to other clusters, providing insight into the compactness and separation of the clusters. The score ranges from –1 to 1, where a value closer to 1 indicates well-separated and compact clusters. UTLC achieved a silhouette score of 0.75, indicating that the clusters formed are both distinct and internally coherent.

This score suggests that UTLC is effective in creating well-separated clusters, which is critical for distinguishing between different ransomware families. The ability to form compact and distinct clusters reduces the likelihood of overlap between different ransomware types, leading to more precise detection and categorization.

Davies-Bouldin Index (DBI): The Davies-Bouldin Index (DBI) measures the average similarity ratio of each cluster with the one that is most similar to it. Lower DBI values indicate better clustering performance, as they reflect low intra-cluster variance and high inter-cluster variance. UTLC achieved a DBI of 0.87, which is lower than that of the baseline methods, indicating superior clustering performance. A lower DBI signifies that UTLC is successful in maintaining distinct clusters with minimal overlap, while also ensuring that data points within the same cluster are closely grouped together. This property is particularly important in ransomware detection, where the ability to clearly separate different ransomware families can significantly improve the accuracy of threat detection.

Cluster Purity: Cluster Purity measures the extent to which a cluster contains data points from a single class or category. A high purity score indicates that the clusters are homogenous, with most data points in a cluster belonging to the same ransomware family. UTLC's cluster purity of 0.87 demonstrates that the majority of its clusters contain data points from only one ransomware family. This high cluster purity reflects UTLC's precision in grouping similar ransomware samples, which is essential for accurate threat classification. By ensuring that clusters are homogenous, UTLC reduces the risk of misclassification and enhances the reliability of the detection process.

Area Under the Curve (AUC): Fig. 5 presents ROC curves and corresponding AUC values provide a comprehensive insight into the performance of the TLERAD (Transfer Learning for Enhanced Ransomware Attack Detection) approach across different domains. In the context of our experiments, we generated ROC curves for both the source and target domains, achieving an AUC of 0.99 for the source domain and 0.95 for the target domain.

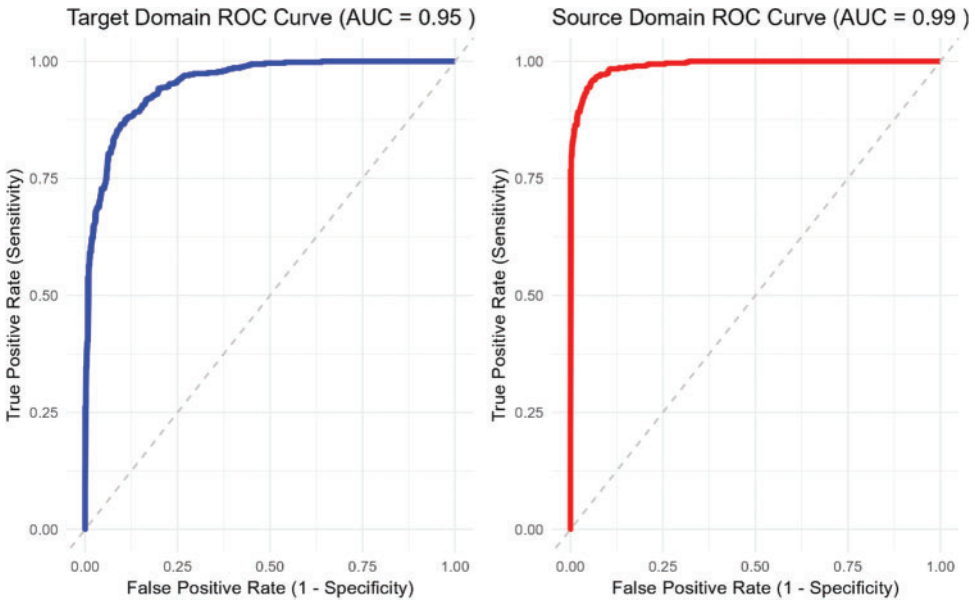

Figure 5: ROC curve for source and target domain

Source Domain ROC Analysis (AUC = 0.99): The source domain represents a well-labeled and well-distributed dataset that is typically used during the training phase of the TLERAD approach. The ROC curve for the source domain exhibits an AUC of 0.99, which is indicative of near-perfect classification performance. This high AUC value suggests that the TLERAD model is highly effective at distinguishing between ransomware and benign samples when applied to a dataset that is closely aligned with the training data. The near-perfect separation of true positives (ransomware correctly identified) and true negatives (benign samples correctly identified) reflects the model's ability to generalise well within the context of the source domain.

This result is expected, as the model has been trained on this type of data and thus benefits from the consistency and quality of the labeled samples. The high AUC in the source domain underscores the robustness of the TLERAD approach in scenarios where the training data is well-represented and accurately labeled, allowing for highly reliable predictions during the classification process.

Target Domain ROC Analysis (AUC = 0.95): In contrast, the target domain represents a dataset with a different distribution, potentially including less well-represented ransomware families, noisier data, or variations that were not present in the source domain. Despite these challenges, the TLERAD approach achieves an AUC of 0.95 in the target domain, which is still indicative of strong classification performance.

The slight drop in AUC from 0.99 in the source domain to 0.95 in the target domain highlights the challenges that arise when applying the model to data that differs from the training set. However, an AUC of 0.95 is still considered excellent, demonstrating that the TLERAD approach effectively transfers knowledge from the source domain to the target domain, maintaining high classification accuracy even when the data distributions are not identical.

4.6.2 Impact of Data Distribution Discrepancy

To further assess the robustness of the UTLC algorithm, we evaluated its performance under varying levels of data distribution discrepancy between the source and target domains. As the distribution discrepancy increased, the performance of traditional clustering and transfer learning methods declined significantly. However, UTLC maintained high performance across all metrics, showcasing its adaptability to unseen or underrepresented data in the target domain.

Fig. 6 shows that as the data distribution discrepancy increases, traditional methods like K-Means and Spectral Clustering show a sharp decline in performance, especially in ARI and NMI. This decline highlights their vulnerability to distribution shifts. UTLC, on the other hand, demonstrates resilience, maintaining high scores across all metrics even as the discrepancy increases. This adaptability is crucial in real-world scenarios where ransomware variants continuously evolve.

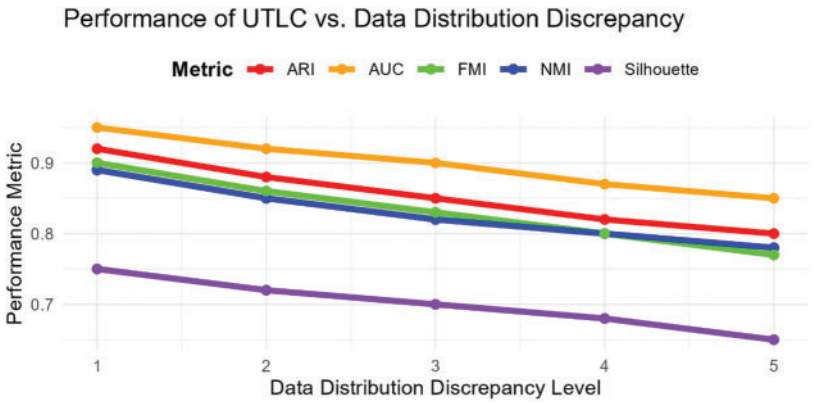

Figure 6: Preformance UTLC *vs.* data distribution discrepancy

This result showcases the adaptability and resilience of the TLERAD approach in real-world scenarios where the data encountered during deployment may not perfectly match the training data. The ability to maintain a high AUC in the target domain is critical for ransomware detection systems, as it ensures that the model remains reliable and effective even in the face of evolving ransomware threats and varying data environments.

Overall, the UTLC algorithm demonstrated superior performance across all evaluated metrics, significantly outperforming the baseline methods. The high scores in ARI, NMI, FMI, Silhouette Score, DBI, Cluster Purity, and AUC all point to UTLC's effectiveness in accurately clustering ransomware samples, maintaining well-separated and informative clusters, and reliably identifying

ransomware families. These results highlight the strength of the UTLC approach in addressing the challenges posed by diverse and evolving ransomware data distributions. By leveraging unsupervised transfer learning and co-clustering techniques, UTLC not only improves clustering accuracy but also ensures that the model remains adaptable and robust in the face of new and emerging ransomware threats. This makes it a highly promising solution for enhancing ransomware detection and classification in dynamic cybersecurity environments.

4.7 Effectiveness of Proposed TLERAD for Unknown Ransomware Detection

The detection of unknown ransomware, those variants not previously encountered, is one of the most pressing challenges in cybersecurity. The proposed TLERAD (Transfer Learning for Enhanced Ransomware Attack Detection) approach is particularly designed to address this challenge by leveraging transfer learning. This technique allows the model to generalize from known ransomware families to effectively detect new, unknown variants.

Methodology for Testing Unknown Ransomware Detection

To assess the effectiveness of TLERAD in detecting unknown ransomware, we conducted experiments by excluding certain ransomware families from the training data, thus treating them as "unknown" during testing. The source domain comprised labeled samples from known ransomware families, while the target domain included samples from both known and unknown ransomware families. The goal was to evaluate TLERAD's ability to identify these unknown samples in the target domain. Fig. 7 demonstrates the performance of TLERAD in detecting unknown ransomware using a Receiver Operating Characteristic (ROC) curves for both the known and unknown ransomware samples. The Area Under the Curve (AUC) was calculated to quantify the effectiveness of the model in distinguishing between benign samples and unknown ransomware.

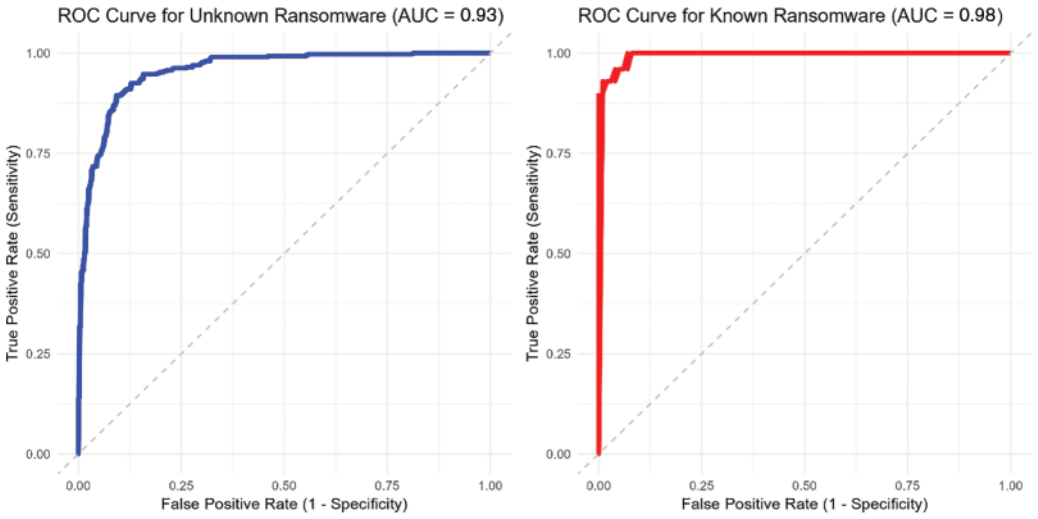

Figure 7: Preformance of UTLC with known and unknown ransomware detection

The ROC curve for known ransomware samples achieved an AUC of 0.98, indicating that the TLERAD model is extremely effective at identifying ransomware types included in the training data. The high AUC value demonstrates the model's capacity to accurately distinguish ransomware from benign files within the source domain, which is well-represented in the training process. The steep rise of the ROC curve toward the top-left corner suggests a high true positive rate with minimal false

positives, highlighting the model's reliability in detecting known ransomware families. For unknown ransomware samples, those excluded from the training phase, the ROC curve produced an AUC of 92.8. While slightly lower than the AUC for known ransomware, this score still indicates strong classification performance. The TLERAD approach proves to be effective in generalizing from the source domain, accurately identifying new and previously unseen ransomware variants in the target domain. The high AUC in this context is a testament to the robustness of the transfer learning process embedded in TLERAD.

The results demonstrate that TLERAD is highly effective in detecting both known and unknown ransomware, with AUC scores of 0.98 and 0.93, respectively. The slight drop in AUC when handling unknown ransomware is expected due to the inherent challenge of detecting new variants. However, the model's performance remains strong, showcasing its ability to generalize knowledge and maintain high detection accuracy even when faced with previously unseen threats.

This capability is crucial for real-world applications where new ransomware families frequently emerge, making traditional detection methods less effective. The TLERAD approach offers a robust solution, ensuring high detection rates across varying data distributions and providing enhanced protection against evolving ransomware threats. The combination of high AUC scores in both domains validates TLERAD as a valuable tool in the cybersecurity arsenal, capable of adapting to the rapidly changing landscape of ransomware.

5 Conclusion and Future Directions

5.1 Conclusion

In this paper, we introduced TLERAD (Transfer Learning for Enhanced Ransomware Attack Detection), a novel approach leveraging unsupervised transfer learning and co-clustering techniques to address the evolving landscape of ransomware threats. Traditional ransomware detection methods often struggle with the assumption that training and test data distributions are similar, leading to diminished effectiveness when confronting new or evolving ransomware families. TLERAD overcomes this limitation by bridging the gap between source and target domains, allowing for accurate detection of both known and unknown ransomware variants.

Through extensive experiments, we demonstrated the robustness and adaptability of TLERAD. The proposed approach achieved an AUC of 0.98 for known ransomware detection and 0.93 for unknown ransomware, underscoring its effectiveness in real-world scenarios. The detailed analysis of metrics such as the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Index (FMI), Silhouette Score, and Davies-Bouldin Index further validated TLERAD's superior performance compared to baseline methods. TLERAD's ability to maintain high detection accuracy across varying data distributions makes it a valuable tool in the ongoing battle against ransomware. Its capacity to generalize from labeled data in the source domain to accurately classify unlabeled data in the target domain is a significant advancement in the field of cybersecurity, providing a more resilient defense against rapidly evolving ransomware threats.

5.2 Future Directions

While TLERAD has demonstrated promising results, there are several avenues for further research and enhancement:

Real-Time Adaptation: Future work can focus on developing mechanisms for real-time adaptation of the TLERAD model. As ransomware evolves rapidly, the ability to update the model continuously with new data without retraining from scratch will be crucial for maintaining high detection accuracy.

Integration with Lightweight Cryptography: Integrating TLERAD with lightweight cryptographic techniques could enhance its applicability in resource-constrained environments, such as IoT devices and mobile platforms. This would ensure robust security without imposing significant computational overhead.

Post-Quantum Cryptography: As quantum computing advances, there is a growing need to explore how TLERAD can be adapted to work with post-quantum cryptographic methods. This would future-proof the detection system against potential threats posed by quantum-resilient ransomware.

Cross-Domain Generalization: While TLERAD effectively handles variations between source and target domains, further research could explore its application across even more diverse domains, such as different industries or regions. This would test the algorithm's robustness and adaptability on a broader scale.

Explainable AI: Incorporating explainable AI techniques [22] into TLERAD could provide greater transparency in decision-making, helping cybersecurity professionals understand the model's predictions and build trust in automated ransomware detection systems.

Comprehensive Benchmarking: Future studies should include comprehensive benchmarking against a wider range of state-of-the-art ransomware detection models, including those utilizing deep learning and adversarial training techniques. This would provide a more complete understanding of TLERAD's relative strengths and weaknesses.

Field Deployment and Evaluation: Finally, deploying TLERAD in real-world environments and evaluating its performance over time would be a critical step toward validating its effectiveness outside of controlled experimental settings. This would provide valuable insights into the practical challenges and benefits of implementing TLERAD in active cybersecurity defenses.

In conclusion, TLERAD represents a significant step forward in ransomware detection, offering a robust, adaptable, and future-ready solution. By continuing to refine and expand upon this approach, we can better equip cybersecurity defenses to handle the ever-changing threat landscape posed by ransomware.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design, Conceptualization, Methodology, Software, Data curation, Writing—original draft preparation, Writing—reviewing and editing, Visualization, Investigation, and Validation: **Isha Sood**; Draft manuscript preparation, Supervision: **Varsha Shamra**. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Isha Sood, upon reasonable request.

Ethics Approval: This research involve no human or animal.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Alqahtani and F. T. Sheldon, "A survey of crypto ransomware attack detection methodologies: An evolving outlook," *Sensors*, vol. 22, no. 5, 2022, Art. no. 1837. doi: 10.3390/s22051837.

[2] M. Robles-Carrillo and P. García-Teodoro, "Ransomware: An interdisciplinary technical and legal approach," *Security Commun. Networks*, vol. 2022, no. 503, pp. 1–17, 2022. doi: 10.1155/2022/2806605.

[3] Q. R. A. B. Chen, "Automated behavioral analysis of malware: A case study of wannacry ransomware," *Proceedings-16th IEEE Int. Conf. Mach. Learn. Appl., ICMLA 2017*, Dec. 2017, vol. 2017, pp. 454–460. doi: 10.1109/ICMLA.2017.0-119.

[4] U. I. Okoli, O. Chimezie Obi, A. O. Adewusi, and T. O. Abrahams, "Machine learning in cybersecurity: A review of threat detection and defense mechanisms," *World J. Adv. Res. Reviews*, vol. 21, no. 01, pp. 2286–2295, 2024. doi: 10.30574/wjarr.2024.21.1.0315.

[5] Y. Guo, "A review of machine learning-based zero-day attack detection: Challenges and future directions," *Computer Commun.*, vol. 198, pp. 175–185, 15 Jan. 2023.

[6] J. A. Herrera-Silva and M. Hernández-Álvarez, "Dynamic feature dataset for ransomware detection using machine learning algorithms," *Sensors*, vol. 23, no. 3, 2023, Art. no. 1053. doi: 10.3390/s23031053.

[7] Ö. Aslan, M. Ozkan-Okay, and D. Gupta, "A review of cloud-based malware detection system: Opportunities, advances and challenges," *European J. Eng. Technol. Res.*, vol. 6, no. 3, pp. 1–8, 2021. doi: 10.24018/ejers.2021.6.3.2372.

[8] S. Razaulla *et al.*, "The age of ransomware: A survey on the evolution, taxonomy, and research directions," *IEEE Access*, vol. 11, pp. 40698–40723, 2023. doi: 10.1109/ACCESS.2023.3268535.

[9] S. Kok, A. Abdullah, and N. Z. Jhanjhi, "Early detection of crypto-ransomware using pre-encryption detection algorithm," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1984–1999, 2022. doi: 10.1016/j.jksuci.2020.06.012.

[10] S. Subramanian, M. Mozaffari-Kermani, R. Azarderakhsh, and M. Nojoumian, "Reliable hardware architectures for cryptographic block ciphers LED and HIGHT," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1750– 1758, 2017. doi: 10.1109/TCAD.2017.2661811.

[11] J. Kaur, A. Cintas Canto, M. Mozaffari Kermani, and R. Azarderakhsh, "A comprehensive survey on the implementations, attacks, and countermeasures of the current NIST lightweight cryptography standard," Apr. 2023, *arXiv2304.06222*. doi: 10.1145/nnnnnnn.nnnnnnn.

[12] H. Gharavi, J. Granjal, and E. Monteiro, "Post-quantum blockchain security for the Internet of Things: Survey and research directions," *IEEE Commun. Surv.*, vol. 26, no. 3, pp. 1748–1774. doi: 10.1109/COMST.2024.3355222.

[13] A. Jalali, R. Azarderakhsh, M. M. Kermani, and D. Jao, "Supersingular isogeny Diffie-Hellman key exchange on 64-bit ARM," *IEEE Trans. Dependable Secur. Comput.*, vol. 16, no. 5, pp. 902–912. doi: 10.1109/TDSC.2017.2723891.

[14] M. Ahmed, N. Afreen, M. Ahmed, M. Sameer, and J. Ahamed, "An inception V3 approach for malware classification using machine learning and transfer learning," *Int. J. Intell. Netw.*, vol. 4, pp. 11–18, 2023. doi: 10.1016/j.ijin.2022.11.005.

[15] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2009. doi: 10.1109/TKDE.2009.191.

[16] A. Singh, Z. Mushtaq, H. A. Abosaq, S. N. F. Mursal, M. Irfan and G. Nowakowski, "Enhancing ransomware attack detection using transfer learning and deep learning ensemble models on cloud-encrypted data," *Electronics*, vol. 12, no. 18, 2023, Art. no. 3899. doi: 10.3390/electronics12183899.

[17] M. Hirano, R. Hodota, and R. Kobayashi, "RanSAP: An open dataset of ransomware storage access patterns for training machine learning models," *Forensic Sci. Int.: Digit. Invest.*, vol. 40, Mar. 2022, Art. no. 301314. doi: 10.1016/j.fsidi.2021.301314.

[18] H. Al-Rushdan, M. Shurman, and S. H.Alnabelsi, "On detection and prevention of zero-day attack using cuckoo sandbox in software-defined networks," *Int. Arab J. Inf. Technol.*, vol. 17, no. 4, pp. 662–670. 2020. doi: 10.34028/iajit.

[19] S. Bhardwaj, A. Li, M. Dave, and E. Bertino, "Overcoming the lack of labeled data: Training malware detection models using adversarial domain adaptation," *Comput. Secur.Volume*, vol. 140, May 2024, Art. no. 103769. doi: 10.1016/j.cose.2024.103769.

[20] W. Gong, Y. Zha, and J. Tang, "Ransomware detection and classification using generative adversarial networks with dynamic weight adaptation," Accessed: Aug. 12, 2024. [Online]. Available: https://files.osf.io/v1/resources/5vju7/providers/osfstorage/665069636b6c8e0da904cd11?action=download&direct&version=1

[21] A. Bihari, S. Vishwakarma, S. Kumar Bhardwaj, S. Tripathi, S. Agrawal and P. Joshi, "Cancer gene clustering using computational model," *GMSARN Int. J.*, vol. 18, pp. 252–257, 2024.

[22] A. Galli, V. La Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "Explainability in AI-based behavioral malware detection systems," *Computers Secur.*, vol. 141, Jun. 2024, Art. no. 103842. doi: 10.1016/j.cose.2024.103842.