Attention and Representation Learning in Byte-Level Digital Forensics: A Survey of Methods, Challenges, and Applications

Teena Mary[1], Sreeja CS[2]
Department of Computer Science, Christ University, Bengaluru, Karnataka, India – 560029[1]
Center for Quantum Technologies and Complex Systems (CQTCS), Christ University, Bengaluru, Karnataka, India - 560029[2]

Abstract—**Byte-level analysis has become an essential capability in digital forensics, enabling content-based investigation when file system metadata, headers, or structural information are unavailable or unreliable. Recent advances in deep learning allow forensic systems to learn discriminative features directly from raw byte streams; however, the growing diversity of representation strategies, architectural designs, and attention mechanisms makes it difficult to assess their relative effectiveness and practical suitability. This study presents a structured survey of representation learning and attention-based approaches for byte-level digital forensic analysis. We examine statistical, embedding-based, image-based, sequential, and hybrid representations, and analyze how architectural choices and attention mechanisms influence performance, robustness, and scalability. Across the literature, hybrid representations combined with lightweight convolutional backbones and selective attention mechanisms consistently provide a favorable balance between accuracy and computational efficiency. The survey also reviews key forensic applications, including file fragment classification, malware and binary analysis, network payload forensics, and encrypted or compressed data triage. In addition, we critically discuss challenges related to distribution shift, dataset bias, adversarial vulnerability, interpretability, and reproducibility, along with practical considerations for deployment in large-scale forensic pipelines. By synthesizing architectural trends, operational constraints, and reliability concerns, this work identifies critical research gaps and provides a structured foundation for the development of robust and trustworthy byte-level forensic learning systems.**

Keywords—Byte-level digital forensics; representation learning; attention mechanisms; file fragment classification; deep learning; forensic robustness

I. Introduction

The exponential growth of digital storage, networked systems, and connected devices has significantly increased both the volume and complexity of digital evidence encountered in modern forensic investigations. In many real-world scenarios, investigators must analyze data that are incomplete, fragmented, partially overwritten, or stripped of file system metadata. Such conditions are common in cases involving file deletion, disk corruption, memory acquisition, network interception, and anti-forensic activities. As a result, traditional metadata-driven forensic techniques often prove insufficient, motivating the adoption of content-based and byte-level analysis methods [1], [2].

Byte-level digital forensics refers to the direct analysis of raw byte sequences without reliance on high-level semantic structures such as file headers, filenames, or file system information. This paradigm underpins several critical forensic and security tasks, including file fragment classification, malware detection, and network payload analysis. Early research in this area relied primarily on handcrafted statistical features, such as byte frequency distributions, entropy measures, and n-gram statistics, to distinguish between different data types [3], [4]. While these approaches demonstrated feasibility, their performance was often limited when applied to high-entropy data, short fragments, or previously unseen file formats.

The emergence of machine learning and deep learning techniques has substantially advanced byte-level forensic analysis by enabling automatic feature extraction directly from raw byte streams. Convolutional neural networks, recurrent architectures, and hybrid models have been successfully applied to learn discriminative patterns from byte sequences without explicit feature engineering [5], [6]. These models have shown improved generalization across diverse file types and forensic conditions, particularly when trained on large-scale benchmark datasets designed to reflect realistic fragmentation scenarios.

Despite these advances, learning from raw byte streams remains fundamentally different from processing structured data such as images or natural language. Byte-level data lack semantic meaning, exhibit high variability, and often display entropy characteristics that closely resemble random noise. Consequently, deep learning models trained on byte sequences are prone to overfitting local statistical artifacts and are highly sensitive to minor perturbations in the input data [7]. These properties pose significant challenges for representation learning, robustness, and interpretability in forensic applications.

In recent years, representation learning and attention mechanisms have emerged as promising tools for addressing these challenges. Representation learning aims to transform raw byte sequences into feature spaces that capture informative statistical and structural regularities, while attention mechanisms enable models to dynamically emphasize discriminative regions or features within otherwise noisy inputs. Attention-based architectures have demonstrated notable success in related security domains, particularly malware classification and binary analysis, by improving both classification accuracy and model efficiency [8], [9]. However, a systematic understanding of how different representation strategies and attention mechanisms operate across byte-level forensic tasks remains lacking.

Existing surveys in digital forensics have primarily focused either on classical file fragment classification techniques or on adversarial vulnerabilities of machine learning models. In contrast, this survey adopts a broader architectural perspective, centering on how byte-level representations are learned, how attention mechanisms are applied, and why these design choices matter for forensic reliability. By synthesizing research across multiple forensic tasks and model families, this study aims to provide a unified foundation for designing effective and trustworthy byte-level forensic learning systems.

However, existing literature lacks a unified analysis that jointly examines representation learning strategies, architectural design choices, attention mechanisms, and their implications for robustness, scalability, and operational deployment in real-world forensic environments. Limited work has systematically connected model design decisions with practical challenges such as distribution shift, high-entropy data, computational constraints, and reliability requirements. This gap motivates the need for a structured synthesis that integrates architectural, methodological, and operational perspectives.

The main contributions of this survey are as follows:

1) A structured review of representation learning strategies for raw byte streams in digital forensics.

2) A comparative analysis of deep learning architectures applied to byte-level forensic tasks.

3) An in-depth examination of attention mechanisms and their role in improving discriminability and robustness.

4) A discussion of key challenges related to reliability, robustness, and trustworthiness in byte-level learning.

5) Identification of open research directions to guide future developments in forensic intelligence systems.

While recent advances in deep learning and attention-based models have significantly improved the effectiveness of byte-level forensic analysis, the design of reliable and robust learning systems remains tightly coupled to the intrinsic properties of byte-stream data. Unlike structured domains such as text or images, raw byte sequences exhibit unique statistical, structural, and entropy-related characteristics that directly influence representation learning, model generalization, and susceptibility to error. A clear understanding of these underlying data properties is therefore essential before examining representation strategies and architectural choices. The following section discusses the fundamental characteristics of byte-level forensic data and outlines the key challenges they impose on learning-based forensic systems.

The remainder of this study is organized as follows: Section II describes the survey methodology and literature selection process. Section III discusses the fundamental characteristics of byte-level forensic data. Section IV reviews representation learning strategies for raw byte streams. Section V examines deep learning architectures used in byte-level forensic analysis, followed by Section VI, which presents attention mechanisms and their role in improving model performance and robustness. Section VII summarizes key application domains, while Section VIII discusses challenges related to robustness, reliability, and trustworthiness. Section IX outlines open research challenges and future directions, and Section X concludes the survey.

II. Survey Methodology

This survey was conducted using a structured literature review approach to ensure comprehensive coverage of research related to byte-level learning for digital forensic applications. Relevant publications were identified through major academic databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar. The search covered the period from 2003 to 2025 to capture both foundational work on file fragment classification and recent advances in deep learning, representation learning, and attention-based architectures.

Search queries included combinations of keywords such as file fragment classification, byte-level analysis, raw byte deep learning, binary classification, representation learning, attention mechanisms, malware raw byte analysis, and encrypted or network payload classification. Retrieved studies were screened based on title and abstract, followed by full-text assessment to ensure relevance to content-based analysis of raw byte data.

Studies were included if they proposed methods for analyzing raw byte streams, investigated deep learning architectures applicable to binary or fragment data, examined representation learning or attention mechanisms for byte-level inputs, or discussed reliability aspects such as distribution shift, adversarial robustness, interpretability, or reproducibility in security or forensic contexts. Works focused solely on metadata-based analysis, signature-based techniques, or unrelated domains without transferable methodological relevance were excluded.

Since research specifically dedicated to file fragment classification remains relatively limited, the scope of this survey was expanded to include closely related byte-level security domains such as malware detection, binary analysis, and encrypted traffic classification. These domains share key characteristics with forensic fragment analysis, including high-entropy inputs, lack of semantic structure, large input sizes, and sensitivity to small perturbations. Insights from these areas were incorporated when they provided transferable principles for representation design, architectural choices, attention mechanisms, or robustness.

In total, the final corpus comprises foundational statistical methods, modern deep learning models, hybrid representation strategies, attention-based architectures, and studies addressing reliability and operational considerations from 60 papers. The selected literature was organized and analyzed according to representation strategy, model architecture, attention mechanism, application domain, and reported limitations to support the comparative synthesis presented in this survey.

III. Characteristics of Byte-Level Forensic Data

As established in the previous section, the effectiveness of learning-based forensic systems is fundamentally constrained by the nature of the data they operate on. Before examining representation learning strategies and model architectures, it is therefore necessary to understand the intrinsic properties of byte-level forensic data that shape both model behavior and performance. Fig. 1 illustrates the shift from traditional metadata-dependent forensic methods to learning-based approaches that operate directly on raw byte streams under conditions of fragmentation, corruption, or missing structural information.

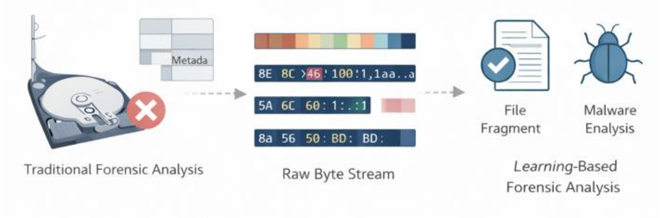

Fig. 1. Conceptual illustration of byte-level digital forensic analysis.

Byte-level forensic analysis operates directly on raw sequences of bytes extracted from storage sectors, memory pages, or network payloads, typically in the absence of file system metadata or semantic context. Unlike structured domains such as text or images, byte streams lack natural token boundaries, spatial coherence, or human-interpretable meaning. Each byte represents only a numerical value between 0 and 255, and any higher-level structure must be inferred implicitly through statistical regularities across the sequence.

One of the most significant characteristics of byte-level forensic data is high-entropy, particularly for fragments originating from compressed, encrypted, or proprietary formats. In such cases, byte distributions often approximate uniform randomness, making class-discriminative patterns weak and difficult to isolate. Recent empirical studies on large-scale fragment benchmarks have shown that even deep learning models experience notable performance degradation as entropy increases, especially when operating on short, fixed-size fragments typical of forensic pipelines [10].

Fragmentation further amplifies this challenge. Forensic tools commonly process data in fixed-size blocks - often aligned with disk sectors rather than logical file boundaries - resulting in fragments that contain only partial structural information. Fragmentation disrupts long-range dependencies and eliminates global cues such as headers, trailers, and format-specific markers. Modern benchmark datasets intentionally remove header information to prevent trivial classification, but this design choice also forces models to rely on subtle local patterns that may not be consistently present across fragments [6].

Another defining property of byte-level forensic data is the absence of semantic interpretability. Unlike words in natural language or pixels in images, individual bytes do not convey meaning that is interpretable by humans. Consequently, learning-based systems must rely almost exclusively on statistical correlations and positional relationships. Recent analyses of byte-level neural networks suggest that many models learn highly localized, distribution-sensitive features, which can lead to overfitting and reduced robustness under realistic forensic conditions involving noise, corruption, or partial overwrites [9].

Byte-level forensic data are also inherently heterogeneous. Different file types, encodings, and storage contexts exhibit distinct statistical behaviors, yet these differences may be subtle or partially obscured at the fragment level. Cross-dataset evaluations have demonstrated that models trained under specific fragmentation or acquisition settings often generalize poorly when exposed to fragments generated under different conditions, highlighting persistent challenges related to dataset bias and distribution shift [11].

Finally, byte-level data are particularly sensitive to minor perturbations. Small modifications - such as padding, bit-level corruption, or partial overwriting - can significantly alter the statistical patterns on which learning-based systems rely. This sensitivity has important implications for reliability and trustworthiness, especially in forensic environments where data integrity cannot be guaranteed. These properties motivate the need for representation learning strategies that can capture discriminative information while remaining resilient to noise and variation [12].

Taken together, these characteristics illustrate why naïve feature engineering and direct application of conventional deep learning architectures are often insufficient for byte-level forensic tasks. They also underscore the central role of representation learning in transforming raw byte sequences into features that are both informative and robust. The next section, therefore, examines the principal representation learning approaches that have been proposed for byte-level digital forensics and analyzes how they address - or fail to address - the challenges outlined above.

IV. Representation Learning for Byte Streams

The data characteristics discussed in the previous section make it clear that raw byte sequences are not directly amenable to effective learning without appropriate feature transformation. Representation learning, therefore, becomes the foundation of byte-level digital forensic analysis, as it determines how statistical regularities, structural cues, and contextual dependencies are exposed to downstream models. Over the past decade, research has evolved from handcrafted statistical descriptors toward learned representations that can adapt to heterogeneous file formats and forensic conditions.

A. Statistical Representations

Early representation strategies for byte-level analysis relied on handcrafted statistical features such as byte frequency distributions, Shannon entropy, and fixed-length n-gram statistics. These representations summarize global properties of a fragment and are computationally efficient, which makes them attractive for large-scale forensic workflows. However, recent empirical studies have shown that statistical descriptors alone are often insufficient when fragments are short, highly compressed, or encrypted, as discriminative patterns become weak or indistinguishable from noise [7], [13]. Moreover, such features are sensitive to fragmentation strategies and dataset bias, limiting their generalization across forensic contexts [11].

Despite these limitations, statistical representations remain relevant as complementary signals. Several recent works incorporate entropy measures or frequency-based features alongside learned representations to improve robustness and stabilize training, particularly in low-data or high-entropy regimes [7].

B. Learned Byte Embeddings

To overcome the rigidity of handcrafted features, recent research has increasingly adopted learned byte embeddings. In this paradigm, each byte value is mapped to a dense, trainable vector that captures contextual relationships among neighboring bytes. Byte embeddings are learned jointly with the classification model and enable the network to discover similarities between byte values that frequently co-occur in specific structural or statistical contexts [4].

Large-scale studies have demonstrated that embedding-based representations significantly outperform raw byte encodings and statistical features on benchmark datasets for file fragment classification and binary analysis [5], [6]. Learned embeddings also reduce input dimensionality and improve optimization stability, making them particularly effective when combined with convolutional architectures. Extensions of this idea have explored context-aware embedding objectives and multi-scale embedding schemes, further improving performance on heterogeneous forensic datasets [14].

C. Image-Based Byte Representations

Another widely explored representation strategy reshapes byte sequences into two-dimensional arrays and interprets them as grayscale images. This approach enables the direct application of convolutional neural networks originally developed for computer vision. Image-based representations have shown competitive results in several byte-level tasks, including malware classification and file fragment identification, due to the ability of CNNs to capture local spatial patterns efficiently [15], [16].

However, the imposed spatial structure is artificial and does not reflect inherent relationships in the byte stream. Recent comparative analyses indicate that while image-based encodings can be effective for fragments containing repeated local motifs, they may obscure long-range dependencies and degrade performance when structural regularities are weak or fragmented [17]. As a result, image-based representations are best viewed as a task-dependent alternative rather than a universal solution.

D. Sequential and Context-Aware Representations

Sequential models explicitly preserve byte ordering and aim to capture contextual dependencies across fragments. Early work employed recurrent neural networks and long short-term memory (LSTM) architectures to model byte sequences, but these approaches suffered from scalability and training inefficiencies on long inputs [18]. More recent studies have investigated the use of self-attention and Transformer-based models for byte-level representation learning, demonstrating improved capacity to model long-range dependencies and contextual interactions [10], [19].

Despite their expressive power, Transformer-based approaches remain computationally expensive and are often impractical for high-throughput forensic pipelines. Consequently, recent research favors lightweight sequential components or attention modules integrated into convolutional backbones to balance expressiveness and efficiency.

E. Hybrid and Multi-View Representations

State-of-the-art systems increasingly adopt hybrid representation strategies that combine multiple views of the data. Common combinations include byte embeddings with convolutional feature maps, statistical descriptors with learned representations, or parallel processing of sequential and image-based encodings. Such multi-view designs aim to mitigate the weaknesses of individual representations and improve robustness across diverse file types and acquisition conditions [20], [21].

Empirical evaluations on large fragment corpora indicate that hybrid representations consistently outperform single-view approaches, particularly in cross-dataset and mixed-entropy scenarios [22]. These findings highlight the importance of representation diversity in addressing the heterogeneity inherent in byte-level forensic data. Table I summarizes the principal representation learning strategies used in byte-level digital forensic analysis, highlighting their core characteristics, strengths, limitations, and typical application contexts.

TABLE I. COMPARISON OF REPRESENTATION LEARNING STRATEGIES FOR BYTE-LEVEL FORENSIC DATA.

Representation Strategy	Core Idea	Strengths	Limitations	Typical Applications
Statistical Representations	Global statistics such as byte frequency, entropy, and n-grams	Simple, fast, interpretable and low computational cost	Poor discrimination for high-entropy or short fragments; limited generalization	Baseline fragment analysis; feature augmentation
Learned Byte Embeddings	Trainable dense vectors for byte values learned end-to-end	Captures contextual byte relationships; compact and efficient	Requires sufficient training data; embedding quality task-dependent	File fragment classification; binary analysis
Image-Based Byte Encodings	Reshaping byte streams into 2D grayscale images	Enables reuse of CNN architectures; strong local pattern learning	Artificial spatial structure; weak long-range dependency modeling	Malware classification; visual binary analysis
Sequential Representations	Explicit modeling of byte order using RNNs or Transformers	Preserves ordering; captures contextual dependencies	High computational cost; scalability challenges	Context-aware fragment analysis; sequence modeling

Representation Strategy	Core Idea	Strengths	Limitations	Typical Applications
Hybrid Representations	Combination of embeddings, statistics, and/or image encodings	Improved robustness; better cross-dataset performance	Increased model complexity; higher training cost	Large-scale forensic pipelines; mixed-entropy datasets

In summary, representation learning has progressed from simple statistical descriptors to sophisticated, multi-view encodings that integrate embeddings, convolutional features, and contextual modeling. However, representation choice alone does not guarantee effective learning. How these representations are processed by downstream architectures - and how attention mechanisms selectively emphasize informative patterns - plays a critical role in determining overall system performance. The next section, therefore, examines the deep learning architectures used to consume these representations and analyzes their respective strengths and limitations.

V. Deep Learning Architectures for Byte-Level Forensics

Selecting an architecture for byte-level forensic tasks is a pragmatic balancing act. Models must be expressive enough to capture discriminative signals buried in noisy fragments, yet efficient enough to process large volumes of data in operational pipelines. In practice, three architectural families dominate the literature- convolutional backbones (often compacted for speed), sequence/contextual models (for longer dependencies), and hybrids that combine local processing with global reasoning. Below, we review each family, highlight important architectural innovations that matter for forensic use, and summarize practical trade-offs.

Convolutional neural networks (CNNs) are widely used because they extract local byte patterns with high efficiency. One-dimensional convolutions over byte sequences or two-dimensional convolutions over reshaped byte images both work in practice; which one is preferable depends on the task and dataset. Lightweight convolutional designs, such as those using depthwise-separable convolutions, substantially reduce computation and memory while preserving accuracy, making them attractive for forensic deployments where throughput or CPU inference is a priority [23], [24]. Recent fragment-classification work explicitly adopts these compact convolutional blocks to achieve near-state-of-the-art accuracy at a fraction of the inference cost [17].

A simple but effective architectural addition is channel-wise recalibration. Modules that learn to reweight feature channels allow networks to suppress noisy dimensions and amplify informative ones. These blocks are inexpensive to insert into compact backbones and improve robustness to heterogeneous fragments by dynamically emphasizing channels that correlate with useful format cues [25]. Empirical studies on byte and binary inputs show that channel recalibration helps when fragments contain sparse but discriminative motifs spread across channels.

Reshaping byte streams into 2-D arrays so they can be processed by image CNNs is a popular engineering choice. It permits reuse of mature image architectures and pretraining tricks. On many benchmark datasets, image-style CNNs match or exceed 1-D counterparts when local texture-like signatures exist in the data, but they can fail when the reshaping imposes artifacts that obscure meaningful sequential correlations [16], [26]. Practitioners, therefore, validate image encodings against sequential approaches before committing to a design.

For fragments, where long-range interactions matter, sequential encoders and Transformer-style models are compelling. Transformers use attention to link distant positions and have been adapted for binary and code analysis tasks with success; they excel when pretraining on large corpora is feasible and when tokenization/patching strategies reduce sequence length [27]. Yet vanilla Transformers are memory-intensive for long byte sequences; practical forensic designs thus use patched tokenizations, sparse attention, or operate Transformers on compressed representations learned by a front-end CNN.

A frequently effective pattern is hybrid: a compact convolutional front-end extracts local features and reduces sequence length; a lightweight attention or recurrent head integrates global context; and shallow classifier heads make the final decision. Hybrids balance expressiveness with throughput and are common in large, published fragment datasets and operational systems where both speed and accuracy matter [5], [28]. They also make it easier to fuse auxiliary features - e.g., entropy statistics or byte histograms - improving robustness under mixed-entropy conditions.

Architecture influences not only accuracy and latency but also robustness to distribution shift and adversarial manipulation. Models that rely excessively on narrow local motifs can be fragile to small perturbations; conversely, architectures that integrate multi-scale context and explicit recalibration mechanisms tend to be more stable. Recent work on binary and malware domains shows that architecture-level choices (tokenization, attention window, channel recalibration) materially affect adversarial susceptibility and cross-dataset generalization, underlining the need to evaluate architectures under diverse and realistic fragment generation scenarios [29], [30], [31].

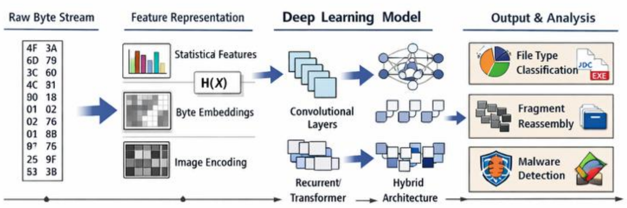

Fig. 2. A generic byte-level deep learning pipeline for digital forensic applications, from raw byte streams to feature representation, model processing, and forensic analysis outputs.

As shown in Fig. 2, the pipeline begins with raw byte streams obtained from storage media or memory, which are converted into suitable feature representations such as statistical features, byte embeddings, or image encodings. These representations are subsequently analyzed by deep learning architectures, including convolutional, recurrent, transformer-based, or hybrid models, to support forensic tasks such as file fragment classification, fragment reassembly, and malware detection.

In practical forensic environments, computational efficiency and scalability are critical considerations, particularly when processing large volumes of data. Byte-level models often operate on long input sequences, which increases memory usage and inference time. Attention-based architectures further amplify this challenge, as self-attention mechanisms scale quadratically with input length, limiting their applicability to large fragments or high-throughput analysis scenarios. To address these constraints, several studies have explored lightweight convolutional architectures and efficient feature extraction strategies that reduce computational overhead while maintaining competitive performance [17], [24]. Hybrid designs that combine convolutional front-end feature extraction with selective or localized attention mechanisms offer a practical trade-off between contextual modeling and efficiency. Such design choices are particularly important for real-world forensic pipelines, where timely processing and resource constraints must be balanced against model accuracy.

VI. Attention Mechanisms in Byte-Level Learning

Architectural advances discussed in the previous section have substantially improved the capacity of deep learning models to process raw byte streams. However, a fundamental limitation remains: byte-level forensic data are inherently noisy, fragmented, and heterogeneous, yet conventional architectures typically treat all parts of the input with equal importance. In realistic forensic scenarios, only a small subset of bytes within a fragment may carry discriminative information, while the remainder consists of padding, overwritten regions, or statistically random content. Attention mechanisms address this mismatch by enabling models to dynamically prioritize informative features and suppress irrelevant signals.

Early applications of attention in byte-level security tasks emerged from malware analysis, where researchers observed that convolutional filters often responded strongly to localized but semantically insignificant byte patterns. Attention mechanisms were introduced to re-weight learned features based on their relevance to the classification objective, leading to consistent improvements in accuracy and stability across datasets [15], [32]. These findings motivated the adoption of attention in file fragment classification and other forensic domains, where similar challenges related to noise and fragmentation are prevalent.

One widely adopted form is channel attention, which operates on feature maps produced by convolutional layers. In byte-level models, each channel typically captures a distinct statistical or structural pattern, such as repetition, local entropy variation, or format-specific motifs. Channel attention modules learn to amplify channels that contribute meaningfully to discrimination while attenuating those dominated by noise. Recent studies demonstrate that incorporating channel attention into compact convolutional backbones improves cross-dataset generalization and reduces sensitivity to fragment length and entropy variation [33], [34]. Importantly, channel attention introduces minimal computational overhead, making it suitable for large-scale forensic pipelines.

Complementing channel-wise reweighting, spatial or positional attention mechanisms focus on identifying informative regions within a byte sequence or its reshaped representation. Spatial attention enables models to localize and emphasize these subsequences, improving resilience to fragmentation and partial overwriting. Empirical evaluations on fragment classification benchmarks show that spatial attention-guided models outperform baseline CNNs, particularly when fragments are generated using non-uniform or adversarial fragmentation strategies [35], [36].

More recently, self-attention and Transformer-inspired architectures have been adapted for byte-level learning. Unlike convolutional or recurrent models, self-attention explicitly models pairwise interactions between positions, allowing long-range dependencies to be captured without sequential processing. In byte-level forensic tasks, self-attention has been applied to embedded byte tokens or compressed patch representations, enabling models to reason about global context across a fragment. Studies report improved performance on complex and high-entropy data, although computational cost remains a significant consideration [37], [38]. To address scalability, several works propose constrained or hierarchical self-attention schemes that limit attention scope while preserving contextual modeling benefits.

Beyond performance gains, attention mechanisms contribute to robustness and interpretability, both of which are critical in forensic settings. By reducing reliance on uniformly distributed statistical noise, attention-equipped models are less susceptible to spurious correlations and dataset-specific artifacts. Furthermore, attention weights can be visualized to provide insights into which regions or features influenced a classification decision. Recent research highlights the value of such visualizations in supporting forensic analysis, model validation, and trustworthiness, particularly when learning-based evidence must be explained or scrutinized [39], [40].

While attention mechanisms enhance feature selection and contextual modeling, their contribution should be interpreted with caution. In many studies, improvements attributed to attention may also be influenced by increased model capacity, deeper architectures, or larger training datasets, making it difficult to isolate the specific impact of the attention component without controlled ablation analysis. Furthermore, attention weights do not necessarily correspond to causal feature importance, and visualizations may create a misleading impression of model interpretability. Prior work in explainable artificial intelligence has emphasized that internal model signals, including attention, should not be treated as faithful explanations without additional validation [54], [59]. Interpretations that are not rigorously evaluated may therefore be unreliable in forensic contexts, where analytical conclusions must be defensible. To support trustworthy deployment, attention-based models should be accompanied by systematic ablation studies and complemented with multiple explanation methods and stability checks, in line with emerging recommendations for rigorous interpretability assessment [55], [60].

Overall, attention mechanisms have increasingly transitioned from auxiliary performance boosters into core architectural components for byte-level forensic learning. Channel, spatial, and self-attention address complementary aspects of the challenges posed by raw byte data, and their integration into hybrid architectures represents a key trend in current research. The next section examines how these attention-enabled models are applied across practical digital forensic tasks, illustrating their effectiveness in real-world investigative scenarios.

VII. Applications in Digital Forensics

The integration of byte-level representation learning and attention mechanisms has expanded the applicability of learning-based methods across several digital forensic domains. These applications share a reliance on raw data analysis under constrained conditions, such as missing metadata, partial data acquisition, or deliberate obfuscation. This section reviews established application areas where byte-level learning has been demonstrably applied and evaluated in the literature.

A. File Fragment Classification

File fragment classification is a long-standing problem in digital forensics, traditionally addressed using statistical features such as byte frequency distributions, entropy measures, and n-grams. Early foundational work demonstrated that fragment classification is feasible even in the absence of file system metadata but also highlighted strong sensitivity to fragmentation strategy and entropy variation [3], [41].

Subsequent studies explored machine learning–based classifiers to reduce reliance on handcrafted features, showing improved discrimination across common file types [4]. While deep learning–based fragment classification remains a relatively narrow research area, convolutional neural networks have been shown to learn discriminative local byte patterns directly from fragment content, outperforming classical approaches under controlled experimental conditions [26]. These results establish the feasibility of byte-level learning for fragment classification, though large-scale and cross-dataset evaluations remain limited.

B. Malware and Binary Artifact Analysis

Byte-level learning has been more extensively studied in malware detection and binary analysis, making this domain a key source of transferable insights for digital forensics. Static malware detection systems operating directly on raw executables have demonstrated that byte-level statistical and structural patterns are sufficient for effective classification without disassembly or dynamic execution [42].

Later work showed that convolutional architectures can automatically learn hierarchical representations of binary structure, while attention mechanisms improve robustness by suppressing noise introduced by padding, packing, or unused code regions [43]. These techniques are directly relevant to forensic analysis tasks such as malware triage, artifact attribution, and family clustering, where raw binaries are often the only available evidence.

C. Network Payload and Traffic Forensics

In network forensics, payload data may be fragmented across packets, partially captured, or encrypted, limiting the effectiveness of protocol-specific parsers. Byte-level deep learning enables content-based traffic analysis without requiring prior knowledge of application protocols. Early deep learning approaches demonstrated that neural networks can classify network traffic directly from raw packet payloads with competitive accuracy [44].

More recent work has emphasized the role of representation learning and attention in improving payload-based traffic classification, particularly under encrypted or obfuscated conditions [45], [46]. These approaches allow models to focus on informative payload segments while ignoring protocol overhead, making them suitable for forensic investigations involving incomplete or noisy traffic captures.

D. Encrypted and Compressed Data Identification

Encrypted and compressed data present inherent challenges due to their high-entropy and lack of semantic structure. While content recovery is infeasible, forensic triage often requires identifying whether data are encrypted, compressed, or random. Prior research has demonstrated that machine learning models trained on byte-level statistics can distinguish between these categories with meaningful accuracy [47].

Recent studies extend this work by applying learning-based classification to encrypted traffic and storage artifacts, supporting forensic prioritization and investigative decision-making [48], [49]. Attention mechanisms contribute by highlighting subtle statistical irregularities that persist across encryption or compression schemes, although performance remains sensitive to dataset composition and entropy levels. To consolidate, Table II provides a summary of representative byte-level forensic tasks along with their associated input characteristics and analytical challenges.

TABLE II. BYTE-LEVEL FORENSIC TASKS AND ASSOCIATED CHALLENGES.

Forensic Task	Input Characteristics	Key Challenges
File fragment classification	Short, high-entropy byte fragments; missing headers and metadata	Loss of global structure; class overlap; limited discriminative cues
Malware binary analysis	Large raw executables with mixed code and data regions	Obfuscation and packing; padding noise; scalability constraints
Network payload analysis	Fragmented packet payloads; partial or encrypted content	Incomplete visibility; protocol noise; payload fragmentation
Encrypted and compressed data triage	Near-random byte distributions with minimal structure	Extremely high-entropy; weak statistical separability

E. Public Datasets and Evaluation Settings

Recent advances in byte-level learning have been supported by the development of several benchmark datasets for file fragment classification and related tasks. However, these datasets differ significantly in terms of fragment generation strategy, header removal, entropy distribution, and evaluation protocols. Such variations complicate direct comparison across studies and contribute to inconsistencies in reported performance.

Large-scale datasets such as FiFTy and ByteRCNN have become commonly used benchmarks for deep learning–based fragment classification. These datasets typically remove file headers and generate fixed-size fragments to prevent trivial classification based on structural signatures. Some studies further stratify fragments based on entropy levels to evaluate performance under realistic forensic conditions involving compressed or encrypted data. Despite these efforts, differences in file type coverage, fragment size, and train–test splitting strategies continue to affect cross-dataset generalization.

Table III summarizes representative publicly reported datasets and experimental settings used in recent byte-level learning studies. The table highlights key factors that influence evaluation outcomes and underscores the need for standardized benchmark protocols to enable fair comparison and reliable assessment of forensic learning systems. The Public column indicates whether datasets or implementation code are explicitly made available through official author resources.

TABLE III. DATASET AND EVALUATION CHARACTERISTICS FOR BYTE-LEVEL FRAGMENT CLASSIFICATION.

Dataset / Study	Type	Public	Header Removed	Key Characteristics
FiFTy [6]	Dataset	Yes (dataset)	Yes	Large-scale benchmark
ByteRCNN [5]	Study	Yes (code)	Yes	CNN–RNN model
ByteNet [35]	Study	Not specified	Yes	Attention-based model
Lightweight CNN [17]	Study	Not specified	Yes	Efficient architecture
Image-based CNN [28]	Study	Not specified	Yes	Grayscale representation

Taken together, these applications demonstrate that byte-level learning supports a range of forensic tasks where traditional feature engineering struggles. However, empirical evidence also reveals limitations related to generalization, robustness, and dataset bias. These concerns motivate a closer examination of reliability and trustworthiness, which are addressed in the following section.

VIII. ROBUSTNESS, RELIABILITY, AND TRUSTWORTHINESS

While byte-level deep learning models demonstrate promising performance across multiple forensic applications, their deployment in real-world investigations raises important concerns related to robustness, reliability, and trustworthiness. Unlike controlled experimental settings, forensic environments are characterized by heterogeneous data sources, imperfect acquisition processes, and potential adversarial manipulation. This section examines key challenges that affect the dependability of byte-level forensic learning systems and reviews established research addressing these issues.

A recurring challenge in byte-level forensic learning is sensitivity to distribution shift. Models are often trained on datasets generated under specific fragmentation strategies, file type distributions, or acquisition conditions. When deployed on data collected under different conditions, performance can degrade significantly. This issue has been widely observed in security and forensic machine learning, where classifiers learn dataset-specific artifacts rather than generalizable patterns [50], [51], leading to significant performance degradation when deployed under different conditions.

Dataset bias is particularly problematic for byte-level models, as subtle differences in compression tools, software versions, or storage media can alter byte distributions. Studies in malware and binary classification demonstrate that high reported accuracy may not translate to robust real-world performance when training and test distributions differ [52]. These findings underscore the need for cross-dataset evaluation and careful dataset construction in forensic research.

Byte-level models are inherently sensitive to small perturbations, such as padding, bit flips, or partial overwriting. Even modifications that preserve semantic equivalence can significantly alter learned representations. Research in adversarial machine learning has shown that deep models operating on raw bytes can be misled by carefully crafted perturbations, raising concerns about reliability in adversarial forensic scenarios [31], [53].

Beyond natural perturbations, byte-level forensic systems may also be vulnerable to deliberate adversarial manipulation. Such threats can be broadly categorized into evasion and poisoning attacks. Evasion attacks occur at inference time and involve modifying input data to mislead the model while preserving its functional or semantic behavior. In byte-level contexts, these manipulations may include byte insertion, padding, reordering, or localized bit modifications that alter statistical patterns without changing the operational characteristics of the file, as demonstrated in adversarial malware studies [29], [31]. Because byte-level models rely heavily on low-level distributional features, even small, localized changes can significantly distort learned representations.

Poisoning attacks target the training process by introducing manipulated or mislabeled samples into the training dataset. In forensic environments, such risks may arise from contaminated data sources, incorrect labeling, or intentionally crafted evidence designed to bias model behavior. Related concerns include evidence manipulation and model inversion, where adversaries attempt to infer sensitive training information or exploit model responses to reconstruct underlying data. These threats highlight the importance of maintaining data integrity and secure training pipelines in forensic machine learning.

Several defensive strategies have been explored in related byte-level security domains. Data augmentation using realistic byte-level transformations can improve robustness to padding, corruption, and minor perturbations, while adversarial training exposes models to perturbed samples during learning to improve resilience. Architectural designs that incorporate multi-scale feature aggregation and attention-based filtering may further reduce sensitivity to localized noise. In addition, techniques such as confidence calibration, out-of-distribution detection, and ensemble-based decision making can help identify uncertain or potentially manipulated inputs.

Although many forensic use cases are not explicitly adversarial, the presence of corrupted or partially overwritten data can produce similar effects. Robustness to such perturbations is therefore essential for trustworthy forensic deployment. Existing work suggests that architectural choices, regularization strategies, and feature aggregation mechanisms influence sensitivity, but no universally robust solution has yet emerged.

Another dimension of reliability concerns generalization across diverse file types and fragment sizes. Byte-level forensic systems are often evaluated on fixed fragment lengths and limited file categories, which may not reflect operational diversity. Prior research indicates that models trained on specific fragment sizes can exhibit degraded performance when applied to shorter or longer fragments [3].

Similarly, generalization across file formats remains challenging, particularly for formats with similar statistical properties. These limitations highlight the importance of evaluating models under varied experimental conditions and avoiding overly narrow performance claims in forensic contexts.

Trustworthiness in forensic applications extends beyond accuracy. Investigators, analysts, and legal stakeholders must be able to understand and justify how conclusions are reached. Deep learning models trained on raw bytes are often criticized for their lack of interpretability, which can hinder their acceptance as forensic tools.

Recent work in explainable artificial intelligence (XAI) emphasizes the need for interpretable models in security-sensitive domains. Techniques such as saliency mapping, feature attribution, and attention visualization provide partial insight into model behavior, but their forensic validity remains an open question [54], [55]. Without careful interpretation, explanations may be misleading or unstable, limiting their usefulness as evidential support.

Reproducibility is a foundational requirement for trust in forensic methods. However, byte-level forensic learning studies often rely on proprietary datasets, undocumented preprocessing steps, or unavailable code, making independent verification difficult. Surveys in digital forensics have repeatedly highlighted reproducibility as a systemic challenge that undermines confidence in experimental results [56].

Transparent reporting of dataset composition, fragmentation strategies, evaluation protocols, and limitations is therefore essential. Without such transparency, even technically sound models may be unsuitable for forensic adoption.

In summary, while byte-level deep learning offers powerful tools for digital forensics, its reliability and trustworthiness cannot be assumed. Sensitivity to distribution shift, noise, and dataset bias, along with challenges related to interpretability and reproducibility, must be carefully addressed. These concerns motivate ongoing research into robust architectures, standardized benchmarks, and explainable learning frameworks. Section IX of this survey discusses open challenges and future research directions that must be addressed to advance byte-level forensic learning toward practical, trustworthy deployment.

IX. Open Challenges and Future Directions

Despite notable advances in byte-level representation learning and attention-based architectures, several unresolved challenges continue to limit their practical adoption in digital forensic investigations. These challenges are not purely algorithmic; rather, they arise from the interaction between data characteristics, model behavior, operational constraints, and evidentiary requirements. Addressing them requires both technical innovation and methodological rigor.

A. Generalization, Robustness, and Real-World Validity

A central challenge in byte-level forensic learning is achieving reliable generalization beyond controlled experimental settings. Many existing studies evaluate models using datasets constructed under specific assumptions about fragmentation strategy, file type distribution, or acquisition process. While such evaluations are necessary for benchmarking, they often fail to capture the variability encountered in real forensic environments. Prior work in digital forensics has repeatedly emphasized that results obtained under narrowly defined conditions may not translate into operational reliability [57].

Robustness to non-ideal data further complicates this issue. Forensic artifacts may be partially overwritten, corrupted, or intentionally manipulated, leading to distribution shifts that degrade model performance. Research in security-focused machine learning shows that models operating on low-level representations are particularly sensitive to subtle perturbations, even when semantic content is preserved [58]. Future research must therefore prioritize systematic robustness evaluation under realistic noise, corruption, and manipulation scenarios, rather than relying solely on clean benchmark performance.

Closely related is the need for standardized, openly available benchmarks that reflect real-world diversity. Without shared datasets and evaluation protocols, it remains difficult to compare approaches objectively or to assess progress meaningfully. The development of representative fragment corpora and transparent evaluation methodologies remains a critical open problem for the field.

B. Interpretability, Scalability, and Forensic Integration

Beyond performance and robustness, the trustworthiness of byte-level forensic learning systems depends heavily on interpretability and operational feasibility. Deep learning models trained on raw byte streams are often opaque, making it difficult for forensic practitioners to understand why a particular classification decision was reached. While attention mechanisms and post hoc explanation techniques offer partial insights, their reliability and evidential value require careful scrutiny. Research in explainable artificial intelligence cautions that visual or attribution-based explanations can be unstable or misleading if not rigorously validated [59].

Scalability presents an additional barrier to adoption. Forensic investigations frequently involve large-scale data processing, such as disk-wide analysis or continuous network monitoring. Although recent architectures have improved efficiency, many attention-based models remain computationally intensive, particularly when applied to long byte sequences. Future work should explore hierarchical processing strategies, lightweight attention mechanisms, and hardware-aware model design to ensure that learning-based methods can be deployed at scale.

Finally, effective integration into forensic workflows remains an open challenge. Automated byte-level analysis should support, rather than replace, human expertise. Forensic methodology research consistently emphasizes the importance of transparency, reproducibility, and analyst oversight in evidentiary processes [56], [60]. Future systems should therefore incorporate confidence estimation, auditability, and human-in-the-loop mechanisms to align learning-based analysis with forensic best practices and legal standards.

In summary, advancing byte-level forensic learning requires more than incremental improvements in model accuracy. Progress depends on addressing foundational challenges related to generalization, robustness, interpretability, scalability, and workflow integration. Meeting these challenges will require collaboration across machine learning, digital forensics, and legal domains and represents a necessary step toward trustworthy and operationally viable byte-level forensic analysis.

X. Conclusion

Byte-level learning has emerged as a powerful paradigm for digital forensic analysis, enabling content-based investigation in scenarios where metadata, structure, or semantic context is unavailable or unreliable. By operating directly on raw byte streams, learning-based systems address long-standing forensic challenges such as file fragment classification, binary artifact analysis, and the triage of encrypted or compressed data. Advances in representation learning and deep architectures have substantially improved the ability of models to extract discriminative signals from noisy and fragmented inputs.

This survey examined how modern forensic systems transform raw bytes into informative representations, how architectural choices influence learning capacity and efficiency, and how attention mechanisms enhance robustness and interpretability. Across the literature, attention-based designs consistently emerge as a unifying theme, allowing models to selectively emphasize informative patterns while suppressing irrelevant or misleading content. These mechanisms are particularly well-suited to forensic data, where discriminative cues are sparse and unevenly distributed.

At the same time, the review highlights important limitations that constrain practical adoption. Sensitivity to dataset bias, distribution shift, and minor perturbations raises concerns about reliability under real-world conditions. The lack of standardized benchmarks and reproducible evaluation protocols further complicates meaningful comparison across studies. Moreover, the interpretability of byte-level models remains an open challenge, particularly when learning-based outputs are expected to support forensic reasoning or legal decision-making.

Moving forward, progress in byte-level forensic learning will depend on addressing these foundational issues rather than pursuing incremental accuracy gains alone. Emphasis on robustness, transparency, and methodological rigor is essential for translating research advances into dependable forensic tools. By aligning representation learning, attention mechanisms, and evaluation practices with the practical realities of digital investigations, future work can help bridge the gap between experimental success and trustworthy forensic deployment. Establishing standardized benchmarks and evaluation practices will be essential for translating research advances into operational forensic capability.

Declaration on the Use of Generative AI

The authors acknowledge the use of generative AI tools for language refinement, grammar correction, and framework creation. All scientific ideas, conceptual structuring, interpretations, and conclusions were developed by the authors. The authors critically reviewed, edited, and validated all AI-assisted content and take full responsibility for the originality, accuracy, and integrity of the manuscript.

References

[1] V. Roussev, “Data fingerprinting with similarity digests,” in Advances in Digital Forensics VI (DigitalForensics 2010), K. P. Chow and S. Shenoi, Eds., IFIP Advances in Information and Communication Technology, vol. 337. Berlin, Heidelberg: Springer, 2010, doi: 10.1007/978-3-642-15506-2_15.

[2] W. C. Calhoun and D. Coles, “Predicting the types of file fragments,” Digit. Invest., vol. 5, pp. S14–S20, 2008, doi: 10.1016/j.diin.2008.05.005.

[3] Karresand and Shahmehri, "File Type Identification of Data Fragments by Their Binary Structure," 2006 IEEE Information Assurance Workshop, West Point, NY, USA, 2006, pp. 140-147, doi: 10.1109/IAW.2006.1652088.

[4] M. McDaniel and M. H. Heydari, "Content based file type detection algorithms," 36th Annual Hawaii International Conference on System Sciences, 2003. Proc. of the, Big Island, HI, USA, 2003, pp. 10 pp.-, doi: 10.1109/HICSS.2003.1174905.

[5] K. Skračić, J. Petrović and P. Pale, "ByteRCNN: Enhancing File Fragment Type Identification With Recurrent and Convolutional Neural Networks," in IEEE Access, vol. 11, pp. 138176-138187, 2023, doi: 10.1109/ACCESS.2023.3340441.

[6] G. Mittal, P. Korus and N. Memon, "FiFTy: Large-Scale File Fragment Type Identification Using Convolutional Neural Networks," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 28-41, 2021, doi: 10.1109/TIFS.2020.3004266.

[7] K. Skračić, J. Petrović, and P. Pale, “Classification of low- and high-entropy file fragments using randomness measures and discrete Fourier transform coefficients,” Vietnam J. Comput. Sci., vol. 10, no. 4, pp. 433–462, 2023, doi: 10.1142/S2196888823500070.

[8] W. K. Al-Ghanem, E. U. H. Qazi, T. Zia, M. H. Faheem, M. Imran, and I. Ahmad, “MAD-ANET: Malware detection using attention-based deep

neural networks," CMES Comput. Model. Eng. Sci., vol. 143, no. 1, pp. 1009–1027, 2025, doi: 10.32604/cmes.2025.058352.

[9] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. K.Nicholas, "Malware Detection by Eating a Whole EXE," in Proc.AAAI Workshop Artif. Intell. Cyber Security, New Orleans, LA, USA, Oct 2017, doi: 10.48550/arxiv.1710.09435.

[10] Y. Wang, W. Liu, K. Wu, K.-H. Yap, and L.-P. Chau, "Intra- and inter-sector contextual information fusion with joint self-attention for file fragment classification," Knowl.-Based Syst., vol. 291, p. 111565, 2024, doi: 10.1016/j.knosys.2024.111565.

[11] V. Roussev and S. L. Garfinkel, "File fragment classification - the case for specialized approaches," in Proc. 4th Int. IEEE Workshop Syst. Approaches Digit. Forensic Eng. (SADFE), 2009, pp. 3–14, doi: 10.1109/SADFE.2009.21.

[12] K. Aryal, M. Gupta, M. Abdelsalam, P. Kunwar and B. Thuraisingham, "A Survey on Adversarial Attacks for Malware Analysis," in IEEE Access, vol. 13, pp. 428-459, 2025, doi: 10.1109/ACCESS.2024.3519524.

[13] A. Bhat, A. Likhite, S. Chavan, and L. Ragha, "File fragment classification using content based analysis," ITM Web Conf., vol. 40, p. 03025, 2021, doi: 10.1051/itmconf/20214003025.

[14] M. E. Haque and M. E. Tozal, "Byte embeddings for file fragment classification," Future Gener. Comput. Syst., vol. 127, pp. 448–461, 2022, doi: 10.1016/j.future.2021.09.019.

[15] M. Basak, D.-W. Kim, M.-M. Han, and G.-Y. Shin, "Attention-based malware detection model by visualizing latent features through dynamic residual kernel network," Sensors, vol. 24, no. 24, Art. no. 7953, 2024, doi: 10.3390/s24247953.

[16] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, "Malware images: Visualization and automatic classification," in Proc. 8th Int. Symp. Vis. Cyber Secur. (VizSec), Pittsburgh, PA, USA, 2011, Art. no. 4, 7 pp., doi: 10.1145/2016904.2016908.

[17] M. Felemban, M. Ghaleb, K. Saaim, S. AlSaleh and A. Almulhem, "File Fragment Type Classification Using Light-Weight Convolutional Neural Networks," in IEEE Access, vol. 12, pp. 157179-157191, 2024, doi: 10.1109/ACCESS.2024.3486180.

[18] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in Proc. 30th Int. Conf. Mach. Learn. (ICML), Atlanta, GA, USA, Jun. 17–19, 2013, vol. 28, no. 3, pp. 1310–1318.

[19] A. Smaili, Y. Zhang, D. E. Mekkaoui, M. A. Midoun, M. Z. Talhaoui, M. Hamidaoui, and W. Kong, "A transformer-based framework for software vulnerability detection using attention-driven convolutional neural networks," Eng. Appl. Artif. Intell., vol. 160, p. 111859, 2025, doi: 10.1016/j.engappai.2025.111859.

[20] H. Huang, Y. Zhou, and F. Jiang, "CLA-BERT: A hybrid model for accurate encrypted traffic classification by combining packet and byte-level features," Mathematics, vol. 13, no. 6, Art. no. 973, 2025, doi: 10.3390/math13060973.

[21] A. Diallo, L. Affognon, C. Diallo and E. C. Ezin, "Deep Learning Based Binary and Multi-class Classification Comparison for Anomaly Detection," 2022 International Conference on Engineering and Emerging Technologies (ICEET), Kuala Lumpur, Malaysia, 2022, pp. 1-6, doi: 10.1109/ICEET56468.2022.10007171.

[22] X. Zhang, H. Huang, D. Zhang, S. Zhuang, S. Han, P. Lai, et al., "Cross-dataset generalization in deep learning," arXiv preprint arXiv:2410.11207, 2024.

[23] E. Prasetyo, R. Purbaningtyas, R. D. Adityo, N. Suciati, and C. Fatichah, "Combining MobileNetV1 and depthwise separable convolution bottleneck with expansion for classifying the freshness of fish eyes," Inf. Process. Agric., vol. 9, no. 4, pp. 485–496, 2022, doi: 10.1016/j.inpa.2022.01.002.

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[25] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[26] Y. Wang, Z. Su, and D. Song, "File fragment type identification with convolutional neural networks," in Proc. Int. Conf. Mach. Learn. Technol. (ICMLT), Jinan, China, 2018, pp. 41–47, doi: 10.1145/3231884.3231889.

[27] K. Wang, M. Wang, Z. Wan, and T. Shen, "Binary transformer based on the alignment and correction of distribution," Sensors, vol. 24, no. 24, Art. no. 8190, 2024, doi: 10.3390/s24248190.

[28] Q. Chen et al., "File Fragment Classification Using Grayscale Image Conversion and Deep Learning in Digital Forensics," 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2018, pp. 140-147, doi: 10.1109/SPW.2018.00029.

[29] K. Li, W. Guo, F. Zhang, and J. Du, "GAMBD: Generating adversarial malware against MalConv," Comput. Secur., vol. 130, p. 103279, 2023, doi: 10.1016/j.cose.2023.103279.

[30] X. Xu, S. Feng, Y. Ye, G. Shen, Z. Su, S. Cheng, G. Tao, Q. Shi, Z. Zhang, and X. Zhang, "Improving binary code similarity transformer models by semantics-driven instruction deemphasis," in Proc. 32nd ACM SIGSOFT Int. Symp. Softw. Test. Anal. (ISSTA), Seattle, WA, USA, 2023, pp. 1106–1118, doi: 10.1145/3597926.3598121.

[31] B. Kolosnjaji et al., "Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 533-537, doi: 10.23919/EUSIPCO.2018.8553214.

[32] F. Deldar and M. Abadi, "Deep learning for zero-day malware detection and classification: A survey," ACM Comput. Surv., vol. 56, no. 2, Art. no. 36, Sep. 2024, doi: 10.1145/3605775.

[33] S. Wu and H. Yamauchi, "A lightweight binarized convolutional block attention module: B-CBAM," J. Adv. Inf. Technol., vol. 16, no. 5, pp. 751–759, 2025, doi: 10.12720/jait.16.5.751-759.

[34] Z. Hou, X. Li, L. Li, J. Yuan and K. Deng, "An End-to-End Raw Bytes Based Malware Classifier via Self-Attention Residual Convolutional Network," 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 2022, pp. 1666-1670, doi: 10.1109/ICCC56324.2022.10065922.

[35] W. Liu, K. Wu, T. Liu, Y. Wang, K. -H. Yap and L. -P. Chau, "ByteNet: Rethinking Multimedia File Fragment Classification Through Visual Perspectives," in IEEE Transactions on Multimedia, vol. 27, pp. 1305-1319, 2025, doi: 10.1109/TMM.2024.3521830.

[36] M. Li, H. Chen, and Z. Cheng, "An attention-guided spatiotemporal graph convolutional network for sleep stage classification," Life, vol. 12, no. 5, Art. no. 622, 2022, doi: 10.3390/life12050622.

[37] T. Sasi, A. H. Lashkari, R. Lu, P. Xiong, and S. Iqbal, "An efficient self attention-based 1D-CNN-LSTM network for IoT attack detection and identification using network traffic," J. Inf. Intell., vol. 3, no. 5, pp. 375–400, 2025, doi: 10.1016/j.jiixd.2024.09.001.

[38] H. Lu, H. Cai, Y. Liang, A. Bianchi and Z. B. Celik, "A Progressive Transformer for Unifying Binary Code Embedding and Knowledge Transfer," 2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), Montreal, QC, Canada, 2025, pp. 383-393, doi: 10.1109/SANER64311.2025.00043.

[39] S. Aribe Jr., "A hybrid deep learning and forensic approach for robust deepfake detection," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 10, 2025, doi: 10.14569/IJACSA.2025.0161028.

[40] F. Ullah, A. Alsirhani, M. M. Alshahrani, A. Alomari, H. Naeem, and S. A. Shah, "Explainable malware detection system using transformers-based transfer learning and multi-model visual representation," Sensors, vol. 22, no. 18, Art. no. 6766, 2022, doi: 10.3390/s22186766.

[41] V. Roussev and G. G. Richard III, "Breaking the performance wall: The case for distributed digital forensics," Digit. Invest., 2004.

[42] B. Anderson, D. Quist, J. Neil, C. Storlie, and T. Lane, "Graph-based malware detection using dynamic analysis," J. Comput. Virol., vol. 7, no. 4, pp. 247–258, 2011, doi: 10.1007/s11416-011-0152-x.

[43] B. Kolosnjaji, A. Zarras, G. Webster, and C. Eckert, "Deep learning for classification of malware system call sequences," in AI 2016: Advances in Artificial Intelligence (AI 2016), B. H. Kang and Q. Bai, Eds., vol. 9992, Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2016, pp. 137–149, doi: 10.1007/978-3-319-50127-7_11.

[44] M. Lotfollahi, M. Jafari Siavoshani, R. Shirali Hossein Zade, and M. Saberian, "Deep packet: A novel approach for encrypted traffic

classification using deep learning," Soft Comput., vol. 24, no. 3, pp. 1999–2012, 2020, doi: 10.1007/s00500-019-04030-2.

[45] X. Lin, G. Xiong, G. Gou, Z. Li, J. Shi, and J. Yu, "ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification," in Proc. ACM Web Conf. (WWW), Virtual Event, Lyon, France, 2022, pp. 633–642, doi: 10.1145/3485447.3512217.

[46] G. Aceto, D. Ciuonzo, A. Montieri and A. Pescapé, "Mobile Encrypted Traffic Classification Using Deep Learning," 2018 Network Traffic Measurement and Analysis Conference (TMA), Vienna, Austria, 2018, pp. 1-8, doi: 10.23919/TMA.2018.8506558.

[47] F. De Gaspari, D. Hitaj, G. Pagnotta, L. De Carli, and L. V. Mancini, "Reliable detection of compressed and encrypted data," Neural Comput. Appl., vol. 34, pp. 20379–20393, 2022, doi: 10.1007/s00521-022-07586-7.

[48] G. Srivastava, M. P. Singh, P. Kumar, and J. P. Singh, "Internet Traffic Classification: A Survey," p. 611, Jun. 2016, doi: 10.1142/9789814704830_0058.

[49] K. Harinath and G. K. Kumar, "An efficient approach for encrypted traffic classification and intrusion detection using packet transformer encoder and CNN," Int. J. Comput. Exp. Sci. Eng., vol. 11, no. 3, 2025, doi: 10.22399/ijcesen.1377.

[50] K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov, "Learning and classification of malware behavior," in Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA 2008), D. Zamboni, Ed. Berlin, Heidelberg: Springer, 2008, vol. 5137, Lecture Notes in Computer Science, pp. 108–125, doi: 10.1007/978-3-540-70542-0_6.

[51] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "DREBIN: Effective and explainable detection of Android malware in your pocket," in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS), 2014.

[52] E. Gandotra, D. Bansal, and S. Sofat, "Malware analysis and classification: A survey," J. Inf. Secur., vol. 5, pp. 56–64, 2014, doi: 10.4236/jis.2014.52006.

[53] H. S. Anderson, A. Kharkar, B. Filar, D. Evans, and P. Roth, "Learning to evade static PE machine learning malware models via reinforcement learning," arXiv preprint arXiv:1801.08917, 2018.

[54] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 51, no. 5, Art. no. 93, Aug. 2019, doi: 10.1145/3236009.

[55] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI Mag., vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[56] G. Horsman, "Framework for Reliable Experimental Design (FRED): A research framework to ensure the dependable interpretation of digital data for digital forensics," Comput. Secur., vol. 73, pp. 294–306, 2018, doi: 10.1016/j.cose.2017.11.009.

[57] D. Lillis, B. A. Becker, T. O'Sullivan, and M. Scanlon, "Current challenges and future research areas for digital forensic investigation," arXiv preprint arXiv:1604.03850, 2016.

[58] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognit., vol. 84, pp. 317–331, 2018, doi: 10.1016/j.patcog.2018.07.023.

[59] Z. C. Lipton, "The mythos of model interpretability," Commun. ACM, vol. 61, no. 10, pp. 36–43, Oct. 2018, doi: 10.1145/3233231.

[60] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.