Deep Learning Based Classification of Security Issues in IoT Devices: A Comprehensive Survey of Architectures, Datasets, and Future Directions

Daisy Merina R

Assistant Professor, Department of Artificial Intelligence and Data Science, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamilnadu, India.

rdaisymerina@gmail.com

ABSTRACT

The exponential growth of Internet of Things deployments has created an unprecedented attack surface characterized by billions of resource-constrained, heterogeneous devices vulnerable to sophisticated cyber threats including Distributed Denial of Service attacks, botnet propagation, malware infiltration, spoofing, and reconnaissance activities that exploit weak authentication, inadequate patching, and limited computational capabilities. Traditional security mechanisms based on signature matching, rule-based intrusion detection, and lightweight heuristics have proven insufficient to address the scale, diversity, and evolving nature of IoT threats, motivating researchers to investigate machine learning and particularly deep learning approaches that can automatically learn complex patterns from network traffic, device telemetry, and behavioral data to classify security incidents with higher accuracy and adaptability than conventional methods. This comprehensive survey synthesizes recent advances in deep learning based classification of security issues in IoT devices, examining how convolutional neural networks extract spatial patterns from flow representations and packet data, how recurrent neural networks including long short-term memory and gated recurrent units model temporal dependencies in sequential attack patterns, how autoencoders enable unsupervised anomaly detection through reconstruction error analysis, how generative adversarial networks facilitate data augmentation and adversarial testing, and how hybrid architectures combining multiple deep learning paradigms achieve superior detection performance compared to single-model approaches. Empirical evidence from studies using benchmark datasets including NSL-KDD, CICIDS2017 and 2018, BoT-IoT, CSE-CIC-IDS2022, and UNB-NB15 demonstrates that deep learning classifiers can achieve detection accuracies exceeding 98 percent for various attack categories, with ensemble methods such as CatBoost and XGBoost reporting accuracies of 98.19 and 98.50 percent respectively on BoT-IoT traffic, while hybrid CNN-LSTM architectures show particular promise for detecting complex multi-stage attacks that exhibit both spatial and temporal characteristics. The survey also identifies persistent challenges including computational complexity and memory requirements that constrain deployment on resource-limited edge devices, high false positive rates in heterogeneous operational environments, class imbalance in training datasets that bias models toward majority classes, vulnerability to adversarial evasion attacks that can manipulate inputs to bypass detection, and limited real-time inference capabilities in fielded systems. Implications for security researchers emphasize the need for adversarially robust models, standardized evaluation protocols, and explainable AI techniques that support incident analysis, while IoT developers should consider hardware-assisted inference, model compression, and federated learning approaches that enable privacy-preserving distributed training without centralizing sensitive data.

KEYWORDS

Internet of Things security, deep learning, intrusion detection, malware classification, convolutional neural networks, long short-term memory, BoT-IoT, federated learning

1. INTRODUCTION

The Internet of Things has fundamentally transformed how physical and digital worlds interact, connecting billions of devices ranging from simple sensors and actuators to sophisticated industrial control systems and autonomous vehicles into vast networks that generate, process, and exchange unprecedented volumes of data. This pervasive connectivity has created immense opportunities for automation, optimization, and intelligence across domains including smart homes, healthcare monitoring, industrial automation, transportation systems, and environmental sensing, yet it has simultaneously introduced severe security challenges that threaten the confidentiality, integrity, and availability of IoT systems and the critical services they support [1]. The security landscape of IoT environments is characterized by several interconnected factors that distinguish it from traditional information technology security contexts and complicate the application of conventional defense mechanisms. First, the massive scale of IoT deployments with projections suggesting tens of billions of connected devices creates an attack surface of unprecedented size where even low-probability vulnerabilities become practically exploitable across the device population [2]. Second, the heterogeneity of IoT devices spanning diverse hardware platforms, operating systems, communication protocols, and application domains makes it difficult to develop uniform security solutions that work across the ecosystem [3]. Third, resource constraints on many IoT devices including limited processing power, memory, battery capacity, and network bandwidth restrict the computational complexity of security mechanisms that can be deployed on endpoints [4]. Fourth, inadequate security practices in IoT device design and deployment including weak default credentials, lack of security updates, unencrypted communications, and absence of authentication create easily exploitable vulnerabilities [5]. Fifth, the long operational lifetimes and physical accessibility of many IoT devices provide attackers with extended opportunities to compromise systems through both cyber and physical attack vectors [6].

The threat landscape facing IoT systems encompasses a wide spectrum of attack types that exploit these vulnerabilities to achieve various malicious objectives. Distributed Denial of Service attacks leverage compromised IoT devices as botnets to flood targets with traffic, exploiting the massive scale and bandwidth of IoT networks to generate attack volumes that overwhelm defensive infrastructure, with notable incidents such as the Mirai botnet demonstrating the devastating potential of IoT-based DDoS attacks [7]. Malware specifically designed for IoT devices including viruses, worms, trojans, and ransomware exploits weak security controls to gain unauthorized access, steal data, disrupt operations, or establish persistent footholds for further attacks [8]. Botnet recruitment and command-and-control activities target vulnerable IoT devices to build armies of compromised endpoints that can be orchestrated for DDoS, spam distribution, cryptocurrency mining, or other malicious purposes [9]. Spoofing and replay attacks manipulate device identities, network addresses, or communication messages to bypass authentication, inject false data, or disrupt legitimate operations [10]. Reconnaissance and scanning activities probe IoT networks to identify vulnerable devices, open ports, default credentials, and exploitable services as precursors to targeted attacks [11]. Data exfiltration and privacy violations exploit inadequate access controls and encryption to steal sensitive information collected by IoT sensors including personal health data, location information, video surveillance, and industrial process parameters [12]. Physical attacks that tamper with device hardware, extract cryptographic keys from memory, or manipulate sensor inputs represent additional threat vectors that are particularly relevant in physically accessible IoT deployments [13].

Traditional security approaches for IoT systems have relied primarily on signature-based intrusion detection, rule-based anomaly detection, and lightweight cryptographic protocols designed to operate within the resource constraints of embedded devices. Signature-based intrusion detection systems maintain databases of known attack patterns and compare observed network traffic or device behaviors against these signatures to identify malicious activities, offering high accuracy for detecting known threats but failing to generalize to novel attacks or polymorphic malware that evades signature matching [14]. Rule-based anomaly detection defines normal system behaviors through manually crafted rules or statistical models and flags deviations from these baselines as potential security incidents, providing some capability to detect unknown attacks but suffering from high false positive rates and difficulty adapting to evolving normal behaviors in dynamic IoT environments [15]. Lightweight cryptographic protocols including constrained versions of TLS, DTLS, and specialized IoT security protocols attempt to provide authentication, confidentiality, and integrity protection within the computational and energy budgets of resource-limited devices, yet many IoT deployments fail to implement even these basic protections due to cost pressures, backward compatibility requirements, or lack of security expertise [16]. Access control mechanisms including role-based and attribute-based access control seek to restrict device interactions and data access according to security policies, but the complexity of configuring and maintaining access control in large-scale, heterogeneous IoT deployments often leads to misconfigurations that create security vulnerabilities [17].

The limitations of traditional security approaches have motivated extensive research into machine learning and particularly deep learning techniques that can automatically learn complex patterns from data to classify security incidents with higher accuracy, adaptability, and generalization capability than rule-based or signature-based methods. Machine learning approaches including decision trees, support vector machines, random forests, and ensemble methods have demonstrated effectiveness for IoT security tasks such as intrusion detection, malware classification, and anomaly detection, offering improvements over traditional methods by learning discriminative features from training data rather than relying on manually crafted rules [18]. Deep learning represents a further evolution that leverages multi-layer neural network architectures to automatically extract hierarchical feature representations from raw or minimally preprocessed data, eliminating the need for manual feature engineering while enabling the modeling of complex nonlinear relationships that characterize sophisticated attacks [19]. The application of deep learning to IoT security classification has accelerated dramatically since 2020, driven by several converging factors including the availability of large-scale labeled datasets for training and evaluation, advances in deep learning architectures and training techniques that improve accuracy and efficiency, increasing computational capabilities of edge devices and cloud platforms that enable deployment of sophisticated models, and growing recognition that conventional security approaches are inadequate to address the scale and sophistication of IoT threats [20].

Deep learning architectures applied to IoT security classification span multiple paradigms, each offering distinct advantages for different aspects of threat detection and analysis. Convolutional neural networks excel at extracting spatial patterns and local correlations from data with grid-like structure such as packet headers represented as images, flow statistics organized in matrices, or spectrograms of network traffic, enabling effective classification of attack types based on their characteristic patterns [21]. Recurrent neural networks including long short-term memory and gated recurrent unit architectures model temporal dependencies in sequential data such as time-series network traffic, system logs, or behavioral traces, capturing the dynamic evolution of attacks that unfold over time and distinguishing between normal and malicious temporal patterns [22]. Autoencoders and variational autoencoders learn compressed representations of input data and reconstruct it, with reconstruction errors serving as anomaly scores that can identify unusual patterns indicative of security incidents without requiring labeled examples of all possible attack types [23]. Generative adversarial networks consisting of generator and discriminator networks in adversarial training can generate synthetic attack samples to address class imbalance in training data, create adversarial examples to test model robustness, and potentially detect evasion attempts by identifying inputs that appear adversarially crafted [24]. Hybrid architectures that combine multiple deep learning paradigms such as CNN-LSTM networks that first extract spatial features through convolutional layers and then model temporal dependencies through recurrent layers often achieve superior performance compared to single-paradigm approaches by leveraging complementary strengths [25].

The motivation for this comprehensive survey stems from the rapid proliferation of research on deep learning for IoT security classification and the need to synthesize findings, identify best practices, and guide future research directions in this critical and evolving field. While several surveys have examined machine learning for IoT security or deep learning for intrusion detection, the landscape continues to evolve rapidly with new architectures, datasets, and evaluation methodologies emerging regularly, creating a need for updated synthesis that reflects the current state of the art [1][2][5]. Furthermore, existing surveys often focus on specific aspects such as particular attack types, specific deep learning architectures, or particular IoT domains, leaving gaps in comprehensive cross-cutting analysis that spans the full range of threats, techniques, and deployment contexts [26]. There is also limited critical analysis of the practical challenges and limitations that constrain the deployment of deep learning security solutions in real-world IoT environments, with most research focusing on benchmark accuracy rather than operational considerations such as computational cost, latency, false positive rates, and adversarial robustness [27]. Additionally, the rapid evolution of the threat landscape with new attack techniques and the corresponding development of defensive approaches create a moving target that requires periodic reassessment to ensure that research efforts remain aligned with practical security needs [28].

The scope of this survey encompasses deep learning approaches for classification of security issues in IoT devices, with particular emphasis on intrusion detection that identifies malicious network traffic or suspicious behaviors, malware classification that categorizes malicious software affecting IoT devices, attack detection that recognizes specific attack types such as DDoS, spoofing, or reconnaissance, and anomaly detection that identifies deviations from normal patterns that may indicate security incidents. The survey covers deep learning architectures including convolutional neural networks, recurrent neural networks and their variants, autoencoders, generative adversarial networks, and hybrid models, examining their application to various security classification tasks. It analyzes commonly used datasets for training and evaluation including NSL-KDD, CICIDS2017 and 2018, BoT-IoT, CSE-CIC-IDS2022, and UNB-NB15, discussing their characteristics, strengths, and limitations. The survey synthesizes empirical findings regarding detection accuracy, computational requirements, real-time performance, and practical deployment considerations, providing comparative analysis across different approaches. It also examines emerging topics including federated learning for privacy-preserving distributed training, explainable AI for interpretable security decisions, adversarial robustness against evasion attacks, and edge-based deployment strategies for resource-constrained environments.

The research objectives of this survey are multifaceted and aim to provide comprehensive coverage that serves multiple stakeholder communities. First, we seek to develop a systematic taxonomy of deep learning architectures for IoT security classification, organizing the diverse landscape of techniques according to their structural characteristics, learning paradigms, and application domains to facilitate understanding and comparison. Second, we aim to synthesize empirical evidence regarding the effectiveness of different deep learning approaches, analyzing reported performance metrics across studies to identify which architectures and techniques demonstrate the most promise for various security classification tasks. Third, we intend to critically evaluate the datasets commonly used for training and evaluating deep learning security classifiers, examining their representativeness, realism, and suitability for assessing model performance in operational IoT environments. Fourth, we seek to identify and analyze the practical challenges and limitations that constrain the deployment of deep learning security solutions in real-world IoT systems, including computational requirements, latency constraints, false positive rates, and adversarial vulnerabilities. Fifth, we aim to highlight emerging research directions and open problems that warrant further investigation, providing guidance for researchers and practitioners working to advance the state of the art in deep learning for IoT security.

The remainder of this paper is organized to provide comprehensive and structured coverage of deep learning based classification of security issues in IoT devices. Section two presents a detailed literature survey that traces the evolution from traditional IoT security approaches to deep learning methods, categorizes the threat landscape and attack types, classifies deep learning architectures according to their characteristics and applications, reviews previous surveys to position the current work, and identifies gaps in existing literature that motivate this synthesis. Section three articulates the research problem statement by clearly defining the security challenges facing IoT systems, formulating specific research questions that guide the survey, and explaining the significance of addressing these questions for advancing both theoretical understanding and practical security capabilities. Section four describes the research methodology including the systematic search strategy used to identify relevant literature, the inclusion and exclusion criteria applied to select papers for detailed analysis, the data extraction procedures used to capture key information, and the limitations of the survey approach. Section five presents comprehensive outcomes and results organized by deep learning architecture, threat category, dataset characteristics, performance metrics, comparative analysis, and deployment considerations, synthesizing findings from the surveyed literature to provide an integrated understanding of the field. Section six concludes with a summary of key findings, implications for different stakeholder groups including researchers, developers, and policymakers, acknowledgment of study limitations, and detailed recommendations for future research directions that can advance the field toward more effective, efficient, and deployable deep learning security solutions for IoT environments.

2. LITERATURE SURVEY

The evolution of security approaches for Internet of Things systems reflects a progression from conventional information technology security mechanisms adapted to IoT constraints, through classical machine learning techniques, to contemporary deep learning methods that leverage hierarchical feature learning and end-to-end training. Early research on IoT security in the 2000s and early 2010s primarily focused on adapting traditional security primitives including lightweight cryptography, secure key management, authentication protocols, and access control to operate within the severe resource constraints of embedded devices [29]. These efforts recognized that standard cryptographic algorithms and security protocols designed for desktop and server environments were often too computationally expensive for resource-limited IoT devices, motivating the development of lightweight alternatives that provided security guarantees with reduced computational overhead [30]. However, cryptographic mechanisms alone proved insufficient to address the full spectrum of IoT security challenges, particularly the detection of attacks that exploit vulnerabilities in device implementations, network protocols, or operational procedures rather than cryptographic weaknesses [31].

The recognition that many IoT security threats manifest as anomalous patterns in network traffic, device behaviors, or system logs motivated the adoption of intrusion detection systems and anomaly detection techniques adapted from traditional IT security contexts. Signature-based intrusion detection systems that match observed patterns against databases of known attack signatures were among the first defensive mechanisms deployed in IoT environments, offering high accuracy for detecting known threats but suffering from fundamental limitations in detecting novel attacks, polymorphic malware, or zero-day exploits that lack established signatures [32]. Anomaly-based intrusion detection systems that model normal system behaviors and flag deviations as potential security incidents provided some capability to detect unknown attacks, but early implementations based on simple statistical models or rule-based heuristics struggled with high false positive rates, difficulty adapting to evolving normal behaviors, and inability to capture complex attack patterns [33]. The limitations of these traditional approaches became increasingly apparent as IoT deployments scaled to millions of devices, attack techniques grew more sophisticated, and the diversity of IoT applications created highly heterogeneous operational environments where one-size-fits-all security solutions proved inadequate [34].

The application of classical machine learning techniques to IoT security classification emerged in the mid-2010s as researchers recognized that supervised learning algorithms could automatically learn discriminative patterns from labeled training data, potentially overcoming the limitations of signature-based and simple anomaly-based approaches. Decision trees and ensemble methods including random forests and gradient boosting machines demonstrated effectiveness for classifying network traffic as normal or malicious based on flow statistics and packet features, offering interpretable models with reasonable computational requirements [35]. Support vector machines with various kernel functions proved capable of learning nonlinear decision boundaries in high-dimensional feature spaces, achieving competitive accuracy for intrusion detection and malware classification tasks [36]. Naive Bayes classifiers provided probabilistic frameworks for classification that could incorporate prior knowledge and handle missing features, though their independence assumptions limited their ability to model complex feature interactions [37]. K-nearest neighbors offered simple instance-based learning that required no explicit training phase, though computational costs for large training sets and sensitivity to feature scaling posed challenges [38]. These classical machine learning approaches represented significant advances over traditional rule-based methods, yet they still required substantial manual feature engineering to extract informative representations from raw data, and their performance plateaued as attack patterns became more complex and subtle [39].

The emergence of deep learning in the late 2010s and its rapid adoption for IoT security classification from 2020 onwards reflects the recognition that multi-layer neural networks can automatically learn hierarchical feature representations from raw or minimally preprocessed data, eliminating the need for manual feature engineering while capturing complex nonlinear relationships that characterize sophisticated attacks. The availability of large-scale labeled datasets for training deep learning models, advances in training techniques including batch normalization, dropout, and adaptive optimization algorithms, and increasing computational capabilities through GPUs and specialized hardware accelerators enabled the practical application of deep neural networks to security classification tasks [40]. Early applications of deep learning to IoT security focused primarily on intrusion detection in network traffic, demonstrating that deep neural networks could achieve higher accuracy than classical machine learning methods by learning complex patterns directly from packet headers, flow statistics, or raw payload data [41]. As the field matured, researchers began exploring specialized deep learning architectures tailored to different aspects of IoT security including convolutional neural networks for spatial pattern extraction, recurrent neural networks for temporal sequence modeling, autoencoders for unsupervised anomaly detection, and hybrid architectures that combined multiple paradigms [42].

The threat landscape facing IoT systems has evolved significantly over the past decade, with attackers developing increasingly sophisticated techniques that exploit the unique characteristics of IoT environments. Distributed Denial of Service attacks have emerged as one of the most prominent threats, with massive botnets assembled from compromised IoT devices generating attack traffic volumes that can overwhelm even well-provisioned defensive infrastructure [43]. The Mirai botnet, which compromised hundreds of thousands of IoT devices using default credentials and launched devastating DDoS attacks in 2016, demonstrated both the vulnerability of IoT devices and their potential as attack platforms [44]. Subsequent botnet variants including Hajime, Reaper, and Mozi have continued to exploit IoT vulnerabilities, incorporating more sophisticated propagation mechanisms, command-and-control protocols, and evasion techniques [45]. Malware specifically designed for IoT devices has proliferated, targeting various platforms including Linux-based embedded systems, real-time operating systems, and custom firmware implementations [46]. These malware families employ diverse infection vectors including exploiting unpatched vulnerabilities, brute-forcing weak credentials, social engineering, and supply chain compromise [47]. Once established, IoT malware can perform various malicious activities including data theft, surveillance, cryptocurrency mining, serving as proxies for other attacks, or destructive actions that brick devices or disrupt operations [48].

Spoofing attacks that manipulate device identities, network addresses, or protocol messages represent another significant threat category, exploiting weak authentication mechanisms and unencrypted communications common in IoT deployments. Address Resolution Protocol spoofing, IP address spoofing, and MAC address spoofing can enable attackers to impersonate legitimate devices, intercept communications, or inject malicious traffic into IoT networks [49]. Man-in-the-middle attacks that intercept and potentially modify communications between IoT devices and backend services exploit the lack of end-to-end encryption in many IoT protocols [50]. Replay attacks that capture and retransmit legitimate messages can bypass authentication or trigger unauthorized actions in systems that lack adequate freshness mechanisms [51]. Reconnaissance and scanning activities that probe IoT networks to identify vulnerable devices, open ports, running services, and exploitable configurations serve as precursors to targeted attacks, with automated scanning tools continuously searching the internet for vulnerable IoT devices [52]. Data exfiltration attacks that steal sensitive information collected or processed by IoT devices pose particular risks in domains such as healthcare, smart homes, and industrial control systems where privacy and confidentiality are critical [53]. Physical attacks including tampering with device hardware, side-channel analysis to extract cryptographic keys, and manipulation of sensor inputs represent additional threat vectors that are especially relevant for IoT devices deployed in physically accessible locations [54].

The classification of deep learning architectures for IoT security reflects the diversity of neural network designs and their suitability for different aspects of threat detection and analysis. Convolutional neural networks have emerged as a dominant architecture for processing data with spatial structure or local correlation patterns, demonstrating particular effectiveness when network traffic or packet data is represented in grid-like formats that preserve spatial relationships [55]. CNNs apply convolutional filters that scan across input data to detect local patterns, with pooling layers that reduce dimensionality while preserving important features, and fully connected layers that perform final classification based on learned feature representations [56]. The hierarchical feature learning in CNNs enables automatic extraction of low-level patterns such as specific byte sequences or protocol fields in early layers, mid-level patterns such as packet structure or flow characteristics in intermediate layers, and high-level semantic concepts such as attack signatures in deeper layers [57]. Applications of CNNs to IoT security include classifying network traffic as normal or malicious based on packet header fields represented as images, detecting malware by converting executable binaries to grayscale images and applying image classification techniques, and identifying attack types from flow statistics organized in matrix formats [58].

Recurrent neural networks including long short-term memory and gated recurrent unit architectures address the temporal dimension of security data by modeling sequential dependencies in time-series network traffic, system logs, or behavioral traces. RNNs maintain internal hidden states that are updated at each time step based on current inputs and previous states, enabling the network to capture temporal patterns and long-range dependencies that characterize many attacks [59]. LSTM networks extend basic RNNs with gating mechanisms that control information flow, addressing the vanishing gradient problem that limits the ability of standard RNNs to learn long-term dependencies [60]. GRU networks provide a simplified gating structure that achieves similar performance to LSTMs with fewer parameters and reduced computational cost [61]. Applications of RNNs to IoT security include detecting botnet command-and-control communications by modeling temporal patterns in network traffic, identifying multi-stage attacks that unfold over time through sequential decision-making, classifying malware based on dynamic behavioral traces captured during execution, and detecting anomalous temporal patterns in sensor data or system logs that may indicate security incidents [62].

Autoencoders represent a class of unsupervised or self-supervised deep learning architectures that learn compressed representations of input data by training encoder networks to map inputs to low-dimensional latent codes and decoder networks to reconstruct the original inputs from these codes. The reconstruction error, measured as the difference between original and reconstructed data, serves as an anomaly score where high reconstruction errors indicate unusual patterns that may represent security threats [63]. Stacked autoencoders with multiple encoding and decoding layers can learn hierarchical representations, while denoising autoencoders trained to reconstruct clean inputs from corrupted versions learn robust features less sensitive to noise [64]. Variational autoencoders extend standard autoencoders by learning probabilistic latent representations and imposing regularization that encourages the latent space to follow a specified distribution, enabling generation of new samples and potentially improving anomaly detection through better-structured latent spaces [65]. Applications of autoencoders to IoT security include unsupervised anomaly detection that identifies unusual network traffic or device behaviors without requiring labeled examples of all possible attack types, dimensionality reduction for preprocessing high-dimensional security data before applying other classifiers, and feature learning that extracts informative representations from raw data for downstream security tasks [66].

Generative adversarial networks consisting of generator and discriminator networks trained in adversarial fashion have found multiple applications in IoT security despite being primarily designed for generative tasks. The generator learns to create synthetic samples that resemble training data, while the discriminator learns to distinguish between real and generated samples, with both networks improving through adversarial training [67]. In security contexts, GANs can generate synthetic attack samples to augment training data and address class imbalance, create realistic benign traffic to improve the training of discriminative models, generate adversarial examples to test the robustness of security classifiers, and potentially detect anomalies by identifying inputs that appear adversarially crafted or inconsistent with the learned data distribution [68]. The application of GANs to IoT security remains an active research area with ongoing exploration of how generative modeling can enhance detection capabilities, improve training data quality, and assess model robustness [69].

Hybrid architectures that combine multiple deep learning paradigms have emerged as particularly promising approaches for IoT security classification, leveraging the complementary strengths of different architectural components to achieve superior performance compared to single-paradigm models. CNN-LSTM networks represent a common hybrid design that first applies convolutional layers to extract spatial features from input data and then feeds these features to LSTM layers that model temporal dependencies, enabling the network to capture both spatial patterns and temporal evolution of attacks [70]. CNN-autoencoder hybrids use convolutional layers in both encoder and decoder paths to learn spatial features while maintaining the unsupervised anomaly detection capability of autoencoders [71]. Attention mechanisms that weight different parts of input sequences or feature maps according to their relevance can be integrated with various architectures to focus on the most informative aspects of security data [72]. Ensemble approaches that combine predictions from multiple deep learning models, potentially with different architectures or trained on different subsets of data, can improve robustness and accuracy by leveraging diverse perspectives [73]. The design of hybrid architectures involves careful consideration of how to effectively integrate different components, balance computational complexity against performance gains, and ensure that the combined model can be trained efficiently [74].

Previous surveys on machine learning and deep learning for IoT security have provided valuable insights while leaving gaps that motivate the current work. Al-Garadi and colleagues conducted a comprehensive survey of machine and deep learning methods for IoT security in 2020, systematically categorizing approaches across multiple security dimensions and providing extensive coverage of classical machine learning alongside emerging deep learning techniques [1]. Their work established a strong foundation for understanding the landscape but predates many recent advances in deep learning architectures, datasets, and deployment strategies that have emerged since 2020. Khan and colleagues surveyed deep learning for intrusion detection and security of IoT in 2022, focusing specifically on deep learning approaches and providing detailed analysis of architectures, datasets, and performance metrics [2]. Their survey offers valuable technical depth on deep learning methods but with less emphasis on practical deployment challenges, adversarial robustness, and emerging topics such as federated learning. Aldhaheri and colleagues reviewed deep learning for cyber threat detection in IoT networks in 2023, examining various deep learning architectures and their applications to different threat categories [5]. Their work provides updated coverage of recent techniques but with limited critical analysis of dataset limitations, cross-dataset generalization, and real-world deployment considerations. Kornaros reviewed hardware-assisted machine learning in resource-constrained IoT environments in 2022, addressing the critical issue of computational constraints and examining how specialized hardware can enable deployment of sophisticated models on edge devices [4]. This survey fills an important gap regarding implementation aspects but focuses less on algorithmic advances and security effectiveness.

The analysis of previous surveys reveals several consistent gaps and limitations that the current work aims to address. First, there is limited integration across the threat taxonomy, deep learning architecture classification, and practical deployment perspectives, with most surveys emphasizing one dimension while treating others peripherally [75]. Second, critical evaluation of datasets used for training and evaluation is often superficial, with insufficient discussion of dataset biases, limitations in representing real-world IoT traffic, and challenges in generalizing across different IoT domains [76]. Third, adversarial robustness and the vulnerability of deep learning security classifiers to evasion attacks receive limited attention despite their critical importance for practical security [77]. Fourth, the gap between reported benchmark accuracy and operational performance in real-world deployments is under-explored, with most research focusing on offline evaluation rather than fielded systems [78]. Fifth, emerging topics including federated learning for privacy-preserving distributed training, explainable AI for interpretable security decisions, and edge-native architectures optimized for resource-constrained deployment warrant more comprehensive coverage [79]. Sixth, standardization of evaluation methodologies, performance metrics, and reporting practices remains limited, complicating cross-study comparison and assessment of relative approach effectiveness [80]. The current survey addresses these gaps by providing integrated coverage across threat types, architectures, and deployment contexts, critical analysis of datasets and evaluation practices, systematic examination of adversarial robustness and practical challenges, and forward-looking discussion of emerging research directions.

3. RESEARCH PROBLEM STATEMENT

The application of deep learning to classification of security issues in IoT devices confronts a complex landscape of technical, operational, and practical challenges that arise from the intersection of sophisticated cyber threats, resource-constrained deployment environments, heterogeneous device ecosystems, and the inherent characteristics of deep learning models themselves. IoT security threats have evolved in sophistication and scale, with attackers leveraging automated tools, exploit kits, and coordinated campaigns to compromise vulnerable devices, assemble botnets, steal data, and disrupt operations, while defensive mechanisms struggle to keep pace with the rate of attack innovation and the diversity of attack vectors [81]. The massive scale of IoT deployments with billions of devices creates an attack surface where even low-probability vulnerabilities become practically exploitable, while the heterogeneity of devices spanning diverse hardware platforms, operating systems, communication protocols, and application domains complicates the development of generalizable security solutions [82]. Resource constraints on many IoT devices including limited processing power, memory capacity, battery life, and network bandwidth restrict the computational complexity of security mechanisms that can be deployed on endpoints, creating tension between the desire for sophisticated deep learning models and the practical limitations of target platforms [83]. The dynamic and evolving nature of IoT threats with attackers continuously developing new techniques to evade detection necessitates adaptive security approaches that can generalize beyond training data, yet achieving robust generalization while avoiding overfitting remains a fundamental challenge in machine learning [84].

The specific characteristics of deep learning models introduce additional considerations that must be addressed when applying them to IoT security classification. Model complexity and the number of parameters in deep neural networks directly affect memory requirements, computational cost, and inference latency, creating challenges for deployment on resource-limited IoT devices or edge gateways [85]. Training data requirements for deep learning can be substantial, particularly for supervised learning approaches that require large volumes of labeled examples, yet obtaining representative labeled datasets for IoT security is challenging due to the diversity of devices and networks, the rarity of some attack types, and the cost of expert labeling [86]. Model interpretability and explainability are limited for complex deep neural networks, creating challenges for security analysts who need to understand why a model flagged particular traffic as malicious, for debugging models that produce false positives or false negatives, and for regulatory compliance in domains where security decisions must be auditable [87]. Adversarial robustness and the vulnerability of deep learning models to carefully crafted perturbations that can cause misclassification raise serious concerns for security applications where attackers have strong incentives to evade detection [88]. Generalization across different IoT deployments, network conditions, and attack variants remains challenging, with models often exhibiting performance degradation when applied to environments that differ from their training distribution [89].

The diversity of IoT security threats further complicates the selection and optimization of deep learning approaches, as different attack categories exhibit distinct characteristics that may favor different architectural choices. Distributed Denial of Service attacks generate high-volume traffic floods with characteristic statistical properties that may be effectively detected through analysis of aggregate flow statistics using convolutional or fully connected networks [90]. Botnet command-and-control communications exhibit temporal patterns and periodic behaviors that may be better captured by recurrent networks that model sequential dependencies [91]. Malware classification requires distinguishing between different malware families based on static features such as binary structure or dynamic features such as behavioral traces, with different deep learning architectures potentially suited to different feature types [92]. Spoofing and replay attacks may manifest as subtle anomalies in protocol fields or timing patterns that require sensitive anomaly detection mechanisms such as autoencoders or one-class classifiers [93]. Multi-stage attacks that involve reconnaissance, exploitation, lateral movement, and data exfiltration unfold over extended time periods and may require models that can capture long-term dependencies and correlate events across multiple time scales [94]. The need to detect diverse attack types with a single model or ensemble of models creates challenges in balancing detection accuracy across categories, managing computational resources, and maintaining acceptable false positive rates [95].

Against this backdrop of challenges and considerations, this survey addresses four primary research questions that structure the investigation and synthesis of the literature. The first research question asks what deep learning architectures and techniques have been applied to classification of security issues in IoT devices, what are their distinguishing characteristics and design rationales, and how do they address specific aspects of the security classification problem. This question motivates a comprehensive taxonomy of deep learning approaches organized by architectural paradigm, learning strategy, and application domain, enabling researchers and practitioners to understand the landscape of available techniques and identify appropriate approaches for specific security challenges [96]. The second research question examines what empirical evidence exists regarding the effectiveness of different deep learning approaches for IoT security classification, considering detection accuracy, false positive and false negative rates, computational requirements, inference latency, and other performance dimensions, and how do these approaches compare against each other and against classical machine learning baselines. This question drives systematic synthesis of reported performance metrics across studies, analysis of experimental conditions and dataset characteristics that influence results, and critical evaluation of the strength of evidence for different approaches [97].

The third research question investigates what datasets are commonly used for training and evaluating deep learning security classifiers for IoT, what are their characteristics, strengths, and limitations, and how well do they represent real-world IoT traffic and attack patterns. This question reflects the recognition that dataset quality and representativeness fundamentally constrain the generalization capability of learned models, motivating detailed analysis of dataset composition, traffic characteristics, attack coverage, labeling quality, and applicability to different IoT domains [98]. The fourth research question identifies what are the key challenges, limitations, and open problems in applying deep learning to classification of security issues in IoT devices, where do current approaches fall short of practical deployment requirements, and what research directions offer the greatest potential for advancing the field toward more effective, efficient, robust, and deployable security solutions. This question ensures the survey provides not only a retrospective synthesis of existing work but also a forward-looking perspective that can guide future research efforts toward addressing the most critical gaps and opportunities [99].

Addressing these research questions provides value to multiple stakeholder communities with different perspectives and priorities. For security researchers in academia and industry, the survey offers comprehensive coverage of the state of the art in deep learning for IoT security classification, identification of open problems and research gaps that warrant further investigation, synthesis of empirical findings that can inform algorithm development and evaluation practices, and contextualization of theoretical advances within practical security requirements and deployment constraints [100]. For IoT system developers and security engineers responsible for implementing defensive mechanisms, the survey provides actionable guidance on selecting appropriate deep learning approaches for specific security challenges, understanding of trade-offs between detection accuracy, computational cost, and operational complexity, practical considerations for deployment including model training, updating, and maintenance, and awareness of potential pitfalls and failure modes that can undermine security effectiveness [101]. For IoT device manufacturers and platform providers, the survey illuminates opportunities to integrate security capabilities into devices and infrastructure, requirements for computational resources and interfaces to support deep learning security applications, and emerging trends that may influence future product development and standardization efforts [102]. For policymakers, standards bodies, and industry stakeholders, the survey offers assessment of technology maturity and readiness for large-scale deployment, identification of barriers to adoption and areas requiring standardization or regulation, understanding of privacy, safety, and ethical implications of deep learning security systems, and perspective on the potential societal and economic impacts of advanced IoT security capabilities [103].

The significance of addressing these research questions extends beyond immediate technical concerns to broader implications for the security and trustworthiness of IoT systems that increasingly underpin critical infrastructure, essential services, and daily life. Effective security classification using deep learning has the potential to detect and respond to threats more quickly and accurately than conventional methods, reducing the window of vulnerability and limiting the damage from successful attacks [104]. Automated threat detection and classification can scale to the massive volumes of IoT devices and traffic that exceed human analytical capacity, enabling security operations centers to focus human expertise on the most critical incidents [105]. Adaptive learning approaches that continuously update models based on new attack patterns can help security systems keep pace with evolving threats, though this must be balanced against the risk of poisoning attacks that manipulate training data [106]. Privacy-preserving techniques such as federated learning that enable collaborative model training without centralizing sensitive data can address concerns about data collection and surveillance while still enabling effective security [107]. By providing a comprehensive, critical synthesis of the current state of deep learning for IoT security classification and identifying priority directions for future work, this survey aims to accelerate progress toward realizing these potential benefits while addressing the challenges and limitations that currently constrain practical deployment.

4. RESEARCH METHODOLOGY

This survey employs a systematic approach to identifying, selecting, and synthesizing relevant literature on deep learning based classification of security issues in IoT devices, following established guidelines for conducting literature reviews in rapidly evolving technical domains while adapting the methodology to accommodate the breadth and interdisciplinary nature of the target literature. The methodology encompasses search strategy and database selection, inclusion and exclusion criteria for paper selection, data extraction and categorization procedures, synthesis and analysis approach, and acknowledgment of methodological limitations. The goal is to provide comprehensive coverage of recent advances while maintaining focus on the most relevant and high-quality contributions that address the core research questions.

The search strategy was designed to capture relevant literature across multiple dimensions including deep learning architectures, security threat types, IoT domains, and evaluation methodologies, using a combination of database searches, forward and backward citation tracking, and consultation of recent surveys to ensure comprehensive coverage. Four major academic databases were systematically searched to ensure broad coverage of published literature across computer science, security, and networking domains. SciSpace provided semantic search capabilities across over 200 million papers with emphasis on recent publications, enabling queries formulated as natural language questions that captured the conceptual scope of the survey [108]. The primary search query used was "What are the recent advances in deep learning based classification of security issues and threats in IoT devices?" supplemented by additional searches targeting specific deep learning architectures, attack types, and application domains. SciSpace Full-Text Search extended coverage by searching within the full text of papers rather than only titles and abstracts, identifying relevant work that might not be captured by metadata-only searches and ensuring that papers discussing deep learning for IoT security in their methodology or results sections were not missed [109]. Google Scholar provided broad coverage including conference proceedings, technical reports, preprints, and gray literature that might not be indexed in traditional academic databases, with keyword searches combining terms such as "deep learning," "IoT security," "threats," "classification," "intrusion detection," "attack detection," and "neural networks" [110]. arXiv covered preprint literature in computer science, machine learning, and security, enabling access to cutting-edge research that may not yet have appeared in peer-reviewed venues, using Boolean searches combining "deep learning," "IoT security," and "classification" with temporal restrictions to focus on recent work [111].

All searches were restricted to publications from 2020 onwards to focus on the recent research era while capturing the transformative impact of advances in deep learning architectures, training techniques, and IoT security challenges during this period. The temporal restriction balances the desire for comprehensive historical coverage against the practical need to manage the scope of the survey and emphasize contemporary approaches most relevant to current and emerging IoT systems. The initial search executed in November 2025 yielded 100 papers from SciSpace, 100 papers from SciSpace Full-Text Search, 20 papers from Google Scholar, and 20 papers from arXiv, for a total initial corpus of 240 papers. These results were then subjected to deduplication to identify and remove papers that appeared in multiple search results, producing a consolidated set of 98 unique papers that formed the primary corpus for detailed analysis.

Inclusion criteria were designed to select papers that make substantive contributions to understanding deep learning approaches for classification of security issues in IoT devices while excluding tangentially related work or papers lacking sufficient technical depth. Papers were included if they met all of the following conditions. First, the publication date fell between January 2020 and November 2025, ensuring focus on recent advances while allowing sufficient time for peer review and publication of work conducted during the target period. Second, the publication type was peer-reviewed journal articles, conference papers, or preprints from reputable venues or archives, providing quality assurance while allowing inclusion of emerging research that may not yet have completed peer review. Third, the content focus explicitly addressed deep learning or neural network approaches applied to security classification in IoT devices, covering topics such as intrusion detection and network traffic classification, malware detection and classification, attack type identification and categorization, anomaly detection for security incidents, or threat intelligence and security analytics. Fourth, the contribution type provided deep learning architectures and methodologies, empirical evaluations and performance comparisons, systematic surveys or reviews, datasets and benchmarking studies, or theoretical analysis and optimization frameworks. Fifth, the paper was written in English and provided sufficient technical detail to extract meaningful information about architectures, datasets, experimental conditions, and results.

Exclusion criteria eliminated papers that were out of scope, lacked technical substance, or provided insufficient information for synthesis. Papers were excluded if they focused solely on IoT security without machine learning or deep learning components, described only classical machine learning approaches without deep learning, presented only conceptual frameworks or position papers without technical implementations or evaluations, were superseded by more comprehensive or recent versions by the same authors, lacked sufficient technical detail about architectures, datasets, or experimental methods, or were purely application-specific case studies without generalizable insights about deep learning approaches. The application of these criteria involved careful judgment in borderline cases, with decisions documented to ensure transparency and consistency across the review process.

Data extraction from included papers followed a structured template designed to capture key information needed to address the research questions and enable synthesis across studies. For each paper, the following elements were systematically recorded. Bibliographic information included authors, title, publication venue, date, and digital object identifier to enable precise citation and retrieval. Deep learning architecture characteristics documented the specific neural network designs employed including architecture family such as CNN, RNN, LSTM, autoencoder, GAN, or hybrid, network topology including number and types of layers, activation functions, and connectivity patterns, model size including number of parameters and memory requirements, and training approach including supervised, unsupervised, semi-supervised, or transfer learning. Security application domain specified the IoT security context including threat types addressed such as DDoS, malware, botnet, spoofing, or general intrusion detection, IoT domains such as smart home, industrial IoT, healthcare, or general networks, and classification granularity such as binary, multi-class, or hierarchical. Dataset characteristics captured information about training and evaluation data including dataset names such as NSL-KDD, CICIDS2017, BoT-IoT, CSE-CIC-IDS2022, dataset size and composition, traffic types and attack categories included, and any preprocessing or feature engineering applied. Performance metrics extracted quantitative results including accuracy, precision, recall, F1-score, false positive rate, false negative rate, detection rate, computational cost and inference time, and comparison against baseline methods. Implementation details documented practical aspects including hardware platforms used for training and inference, software frameworks such as TensorFlow, PyTorch, or Keras, and deployment considerations for edge or cloud environments. Identified limitations and challenges noted issues and constraints acknowledged by the authors including dataset limitations, computational constraints, generalization challenges, and adversarial robustness concerns.

Analysis and synthesis of the extracted data proceeded through multiple stages to organize findings and address the research questions. Categorical mapping developed a comprehensive taxonomy of deep learning architectures for IoT security classification by identifying common architectural patterns and design principles across papers, organizing techniques by learning paradigm, structural characteristics, and application focus, and creating hierarchical categorizations that capture relationships between approaches and their variants [112]. Thematic analysis identified recurring patterns and insights including benefits and advantages of deep learning approaches reported across studies, challenges and limitations that constrain practical deployment, trade-offs between competing objectives such as accuracy versus computational cost, and best practices and design principles emerging from successful implementations [113]. Comparative synthesis examined empirical performance results by aggregating quantitative findings across studies where comparable metrics and conditions permitted, identifying factors that influence relative performance including architecture choices, dataset characteristics, and experimental setups, assessing the strength of evidence for different approaches based on study quality and consistency of findings, and noting inconsistencies or contradictions in reported results that warrant further investigation [114]. Gap analysis identified areas where knowledge is limited or absent by comparing the scope of existing work against the full range of relevant topics, noting methodological limitations that affect interpretation of results such as dataset biases or evaluation practices, and highlighting open problems and research opportunities that warrant future investigation [115].

Quality assessment of included papers considered multiple dimensions of methodological rigor and contribution significance, though formal quality scoring was not applied given the diversity of paper types and research approaches. For empirical studies, assessment considered the appropriateness of experimental design including dataset selection, train-test split, and cross-validation procedures, size and representativeness of datasets used for evaluation, rigor of performance evaluation including statistical significance testing and comparison against baselines, and transparency regarding limitations, potential confounds, and threats to validity. For surveys and reviews, assessment evaluated comprehensiveness of coverage across relevant topics and literature, critical analysis versus purely descriptive summary, identification of gaps and future directions, and currency and relevance to contemporary challenges. For architectural and algorithmic papers, assessment examined novelty of contributions relative to prior work, mathematical rigor and soundness of theoretical analysis, and connection to practical applications and empirical validation. This multidimensional quality assessment informed the relative weight given to different papers in the synthesis while avoiding rigid exclusion based on single quality dimensions.

The methodology has several important limitations that should be considered when interpreting the survey findings. The search strategy, while systematic and comprehensive, may have missed relevant papers published in specialized venues not well-indexed by the selected databases, work published in languages other than English, or very recent work that had not yet been indexed or published at the time of the search. The temporal restriction to 2020 onwards excludes potentially relevant foundational work from earlier periods, though this trade-off was deemed necessary to maintain focus on contemporary approaches and manage the scope of the survey. The inclusion and exclusion criteria involve subjective judgments that other researchers might apply differently, potentially affecting which papers were included in the final corpus and how borderline cases were handled. The data extraction process relied on information provided in the papers themselves, which may be incomplete, inconsistent across papers, or subject to reporting biases that favor positive results and downplay negative findings or limitations. The synthesis and analysis reflect the authors' interpretation and organization of the literature, and alternative taxonomies, analytical frameworks, or emphasis on different aspects might yield different insights or conclusions. The heterogeneity of experimental conditions, datasets, preprocessing approaches, and performance metrics across papers limits the extent to which quantitative results can be directly compared or aggregated, necessitating primarily qualitative synthesis. Publication bias likely results in underrepresentation of negative results, failed approaches, and limitations of successful methods, potentially creating an overly optimistic view of the state of the art. The rapidly evolving nature of both IoT threats and deep learning techniques means that some findings may be superseded by more recent work even during the survey preparation period. These limitations are inherent to survey research in fast-moving interdisciplinary fields and are mitigated through transparent reporting, cautious interpretation of findings, and explicit acknowledgment of uncertainty where appropriate.

5. OUTCOMES AND RESULTS

The synthesis of 98 papers on deep learning based classification of security issues in IoT devices reveals a rapidly maturing field characterized by diverse architectural approaches, extensive empirical validation, and increasing sophistication in addressing the unique challenges of IoT security contexts. This section presents comprehensive findings organized by deep learning architecture taxonomy, security threat categories, dataset characteristics, performance metrics and comparative analysis, real-time deployment considerations, and persistent challenges, drawing on empirical results, expert perspectives, and critical analysis from the surveyed literature to provide an integrated understanding of the current state of the field.

The taxonomy of deep learning architectures for IoT security classification organizes approaches according to their structural characteristics, learning paradigms, and primary application focus, providing a structured framework for understanding the diverse techniques employed in recent research. Convolutional neural networks represent the most widely applied architecture family for IoT security classification, leveraging their ability to extract spatial patterns and local correlations from data with grid-like structure [116]. CNNs applied to network traffic classification typically transform packet headers or flow statistics into two-dimensional representations that preserve spatial relationships, enabling convolutional filters to detect characteristic patterns associated with different attack types [117]. The convolutional layers apply learnable filters that scan across input data, with each filter detecting specific patterns such as particular byte sequences, protocol field combinations, or statistical distributions [118]. Pooling layers reduce spatial dimensions while preserving important features, providing translation invariance and reducing computational requirements for subsequent layers [119]. Fully connected layers at the network output perform final classification based on high-level features extracted by convolutional and pooling layers [120]. Applications of CNNs to IoT security include binary classification of traffic as normal or malicious, multi-class classification of specific attack types such as DDoS, malware, or reconnaissance, and hierarchical classification that first distinguishes attack categories and then identifies specific attack variants within categories [121].

Recurrent neural networks including long short-term memory and gated recurrent unit architectures address the temporal dimension of security data by modeling sequential dependencies in network traffic, system logs, or behavioral traces. RNNs maintain hidden states that are updated at each time step based on current inputs and previous states, enabling the network to capture temporal patterns that unfold over time [122]. LSTM networks extend basic RNNs with gating mechanisms including input gates that control what new information is added to the cell state, forget gates that determine what information is discarded, and output gates that regulate what information is exposed to subsequent layers [123]. These gates enable LSTMs to selectively retain or discard information over long sequences, addressing the vanishing gradient problem that limits standard RNNs to modeling only short-term dependencies [124]. GRU networks simplify the LSTM architecture by combining the forget and input gates into a single update gate and merging the cell state and hidden state, achieving similar performance with fewer parameters and reduced computational cost [125]. Applications of RNNs to IoT security include detecting botnet command-and-control communications by modeling temporal patterns in network flows, identifying multi-stage attacks that exhibit characteristic temporal sequences of activities, classifying malware based on dynamic execution traces captured over time, and detecting anomalous temporal patterns in sensor data or device behaviors that may indicate compromise [126].

Autoencoders and their variants provide unsupervised or self-supervised approaches to security classification that learn compressed representations of input data and use reconstruction errors as anomaly scores. Standard autoencoders consist of encoder networks that map inputs to low-dimensional latent representations and decoder networks that reconstruct inputs from these representations, with the reconstruction error serving as a measure of how unusual an input is compared to the training distribution [127]. Stacked autoencoders with multiple encoding and decoding layers learn hierarchical representations that capture different levels of abstraction, while denoising autoencoders trained to reconstruct clean inputs from corrupted versions learn robust features less sensitive to noise and minor perturbations [128]. Variational autoencoders extend standard autoencoders by learning probabilistic latent representations and imposing regularization that encourages the latent space to follow a specified distribution such as a Gaussian, enabling both reconstruction and generation of new samples [129]. Applications of autoencoders to IoT security include unsupervised anomaly detection that identifies unusual network traffic or device behaviors without requiring labeled examples of all possible attack types, semi-supervised learning that leverages large amounts of unlabeled data alongside limited labeled examples, dimensionality reduction for preprocessing high-dimensional security data before applying supervised classifiers, and feature learning that extracts informative representations from raw data for downstream security tasks [130].

Generative adversarial networks consisting of generator and discriminator networks trained in adversarial fashion have found multiple applications in IoT security despite being primarily designed for generative tasks. The generator learns to create synthetic samples that resemble training data, while the discriminator learns to distinguish between real training samples and generated samples, with both networks improving through adversarial training where the generator tries to fool the discriminator and the discriminator tries to correctly classify samples [131]. In security contexts, GANs can generate synthetic attack samples to augment training data and address class imbalance problems where some attack types are underrepresented, create realistic benign traffic to improve the training of discriminative models, generate adversarial examples to test the robustness of security classifiers against evasion attacks, and potentially detect anomalies by identifying inputs that appear adversarially crafted or inconsistent with the learned data distribution [132]. The application of GANs to IoT security remains an active research area with ongoing exploration of how generative modeling can enhance detection capabilities, improve training data quality, and assess model robustness against adversarial manipulation [133].

Hybrid architectures that combine multiple deep learning paradigms have emerged as particularly promising approaches for IoT security classification, leveraging the complementary strengths of different architectural components to achieve superior performance compared to single-paradigm models. CNN-LSTM networks represent the most common hybrid design, first applying convolutional layers to extract spatial features from input data such as packet headers or flow statistics, and then feeding these features to LSTM layers that model temporal dependencies across sequential time steps [134]. This architecture enables the network to capture both the spatial patterns that characterize individual packets or flows and the temporal evolution of traffic that distinguishes different attack types [135]. CNN-autoencoder hybrids use convolutional layers in both encoder and decoder paths to learn spatial features while maintaining the unsupervised anomaly detection capability of autoencoders, potentially improving reconstruction quality and anomaly detection sensitivity [136]. Attention mechanisms that weight different parts of input sequences or feature maps according to their relevance for the classification task can be integrated with various architectures to focus on the most informative aspects of security data, improving both accuracy and interpretability [137]. Ensemble approaches that combine predictions from multiple deep learning models, potentially with different architectures or trained on different subsets of data, can improve robustness and accuracy by leveraging diverse perspectives and reducing the risk of systematic errors [138].

The categorization of security threats in IoT environments provides context for understanding which deep learning approaches are most appropriate for different attack types and detection scenarios. Distributed Denial of Service attacks generate high-volume traffic floods aimed at overwhelming target systems, with characteristic statistical properties including elevated packet rates, skewed source address distributions, and abnormal protocol distributions that can be detected through analysis of aggregate flow statistics [139]. Deep learning approaches for DDoS detection typically employ CNNs or fully connected networks that classify traffic flows based on statistical features, with reported detection accuracies often exceeding 95 percent on benchmark datasets [140]. Botnet activities including propagation, command-and-control communications, and coordinated attacks exhibit temporal patterns and periodic behaviors that can be captured by RNNs and LSTMs, with studies reporting detection rates above 90 percent for botnet traffic identification [141]. Malware classification distinguishes between different malware families affecting IoT devices based on static features such as binary structure, API calls, or file characteristics, or dynamic features such as behavioral traces captured during execution, with deep learning approaches achieving accuracies above 95 percent on various malware datasets [142]. Spoofing attacks that manipulate device identities, network addresses, or protocol messages may manifest as subtle anomalies that require sensitive detection mechanisms such as autoencoders that can identify deviations from normal patterns [143]. Reconnaissance and scanning activities that probe networks to identify vulnerable devices exhibit characteristic patterns in port scanning sequences, timing distributions, and target selection that can be detected through sequence modeling with RNNs [144].

Dataset characteristics and the availability of high-quality labeled data fundamentally constrain the training and evaluation of deep learning security classifiers for IoT. NSL-KDD represents an improved version of the classic KDD Cup 1999 dataset, removing duplicate records and addressing some of the original dataset's limitations, though it still reflects network traffic from the late 1990s that may not represent contemporary IoT traffic patterns [145]. CICIDS2017 and CICIDS2018 provide more recent network traffic captures including both benign traffic and various attack types such as DDoS, infiltration, and web attacks, with labeled flows and extensive feature sets that have made these datasets widely used for evaluating intrusion detection systems [146]. BoT-IoT specifically focuses on IoT traffic and includes captures from a realistic IoT network testbed with various IoT devices, providing both benign traffic and multiple attack types including DDoS, reconnaissance, and information theft, making it particularly relevant for IoT security research [147]. CSE-CIC-IDS2022 represents one of the most recent comprehensive intrusion detection datasets, including contemporary attack types and modern network traffic characteristics, though its adoption in published research is still growing [148]. UNB-NB15 provides another recent dataset with diverse attack types and detailed flow features, though some researchers have noted concerns about labeling accuracy and traffic realism [149]. The characteristics of these datasets including their size, attack coverage, traffic realism, labeling quality, and representativeness of actual IoT deployments significantly influence the performance and generalizability of models trained on them [150].

Empirical performance results synthesized from the surveyed literature demonstrate that deep learning approaches can achieve high detection accuracies on benchmark datasets, though performance varies significantly depending on the specific architecture, dataset, and experimental conditions. Studies using BoT-IoT traffic report that ensemble methods such as CatBoost and XGBoost can achieve accuracies of 98.19 percent and 98.50 percent respectively under configured feature engineering and sampling pipelines, demonstrating that even classical gradient-boosted models can achieve excellent performance when applied to well-characterized traffic with informative features [3]. Deep learning approaches including CNNs, LSTMs, and hybrid CNN-LSTM architectures typically report accuracies in the range of 95 to 99 percent on various datasets for binary classification of traffic as normal or malicious [151]. Multi-class classification that distinguishes between specific attack types generally shows lower accuracy than binary classification, with reported accuracies typically in the range of 90 to 97 percent depending on the number of classes and their separability [152]. Precision and recall metrics provide additional insights beyond overall accuracy, with many studies reporting precision and recall values above 95 percent for well-represented attack classes, though minority classes in imbalanced datasets often show significantly lower recall [153]. False positive rates, which are critical for practical deployment to avoid alert fatigue, vary widely across studies with reported values ranging from less than 1 percent to over 10 percent depending on the detection threshold and model configuration [154].

Comparative analysis across different deep learning architectures reveals several general patterns, though the heterogeneity of experimental conditions limits definitive conclusions. Hybrid architectures that combine CNNs for spatial feature extraction with LSTMs for temporal modeling generally achieve higher accuracy than single-architecture approaches on datasets where both spatial and temporal patterns are informative [155]. Deep learning approaches typically outperform classical machine learning methods such as decision trees, support vector machines, and random forests by margins of 2 to 10 percent in terms of accuracy, though the advantage is smaller when classical methods are provided with carefully engineered features [156]. Autoencoders for unsupervised anomaly detection show promise for detecting novel attacks not seen during training, though they typically achieve lower accuracy than supervised approaches on known attack types [157]. The computational cost of deep learning models varies significantly, with CNNs generally requiring less computation than RNNs for inference on fixed-size inputs, though RNNs may be more efficient for variable-length sequences [158]. Model size and memory requirements also vary, with some architectures requiring hundreds of megabytes for parameter storage that may be prohibitive for edge deployment, while compressed models can achieve acceptable accuracy with memory footprints in the tens of megabytes [159].

Real-time detection capabilities and deployment considerations represent critical practical concerns that receive limited attention in much of the published research. Most studies report offline evaluation where models are trained on historical data and tested on held-out test sets, with inference times measured on workstation or server hardware that may not reflect the constraints of actual IoT edge devices [160]. The few studies that examine real-time deployment report inference latencies ranging from milliseconds to seconds depending on model complexity and hardware platform, with simpler models achieving sub-millisecond inference on embedded processors while complex models may require tens to hundreds of milliseconds [161]. Hardware acceleration through GPUs, neural processing units, or field-programmable gate arrays can reduce inference time by factors of 10 to 1000 compared to CPU-only execution, though the availability and cost of specialized hardware must be considered [162]. Model compression techniques including pruning, quantization, and knowledge distillation can reduce model size and computational requirements by factors of 5 to 100 with modest accuracy degradation, making sophisticated models more feasible for resource-constrained deployment [163]. Edge computing architectures that partition processing between IoT devices, edge gateways, and cloud servers can balance the trade-offs between local inference for low latency and centralized processing for sophisticated analysis [164].

Persistent challenges and limitations identified across the surveyed literature highlight areas where current approaches fall short of practical requirements and future research is needed. Computational complexity and resource requirements of deep learning models often exceed the capabilities of resource-constrained IoT devices, necessitating either deployment on more capable edge gateways or cloud platforms with associated latency and privacy concerns, or use of simplified models that may sacrifice accuracy [165]. High false positive rates in operational environments with heterogeneous traffic and diverse normal behaviors can lead to alert fatigue and reduced trust in automated detection systems, requiring careful threshold tuning and potentially human-in-the-loop verification [166]. Class imbalance in training datasets where benign traffic vastly outnumbers attack traffic and some attack types are rare can bias models toward majority classes, resulting in poor detection of minority attack types that may be of high security concern [167]. Adversarial robustness and vulnerability to evasion attacks where attackers craft inputs to bypass detection represent serious concerns for security applications, yet most research does not evaluate models against adversarial examples or adaptive attackers [168]. Generalization across different IoT deployments, network conditions, and device types remains challenging, with models often exhibiting significant performance degradation when applied to environments that differ from their training distribution [169]. Dataset limitations including lack of realism in synthetic traffic, labeling errors, limited attack coverage, and rapid obsolescence as attack techniques evolve constrain the ability to train models that perform well in real-world operational environments [170]. Lack of interpretability and explainability in complex deep learning models makes it difficult for security analysts to understand why particular traffic was flagged as malicious, complicating incident response and debugging of false positives [171]. Privacy concerns arise when training data contains sensitive information about network activities or device behaviors, motivating research into privacy-preserving learning techniques such as federated learning and differential privacy [172].

6. CONCLUSION

This comprehensive survey has synthesized recent advances in deep learning based classification of security issues in IoT devices, covering architectural approaches, security threat categories, datasets and evaluation methodologies, empirical performance results, deployment considerations, and persistent challenges based on analysis of 98 papers published between 2020 and 2025. The findings reveal a field characterized by rapid innovation in deep learning architectures, extensive empirical validation on benchmark datasets, and increasing recognition of practical challenges that must be addressed to transition from laboratory demonstrations to operational deployments.

The key findings from this survey can be organized into several overarching themes that characterize the current state of the field. First, deep learning approaches including convolutional neural networks, recurrent neural networks, autoencoders, and hybrid architectures have demonstrated superior performance compared to traditional signature-based and rule-based intrusion detection systems, achieving detection accuracies exceeding 95 percent and often reaching 98 to 99 percent on benchmark datasets for binary classification of traffic as normal or malicious [3][151]. Second, hybrid architectures that combine multiple deep learning paradigms such as CNN-LSTM networks that extract both spatial and temporal features show particular promise for detecting complex multi-stage attacks, often outperforming single-architecture approaches by margins of 2 to 5 percent in terms of accuracy [155]. Third, the availability of realistic labeled datasets including CICIDS2017/2018, BoT-IoT, and CSE-CIC-IDS2022 has enabled systematic evaluation and comparison of different approaches, though concerns remain about dataset representativeness, labeling quality, and generalization to real-world IoT traffic [147][150]. Fourth, computational requirements and resource constraints pose significant challenges for deploying sophisticated deep learning models on resource-limited IoT edge devices, motivating research into model compression, hardware acceleration, and edge-cloud architectures [165]. Fifth, adversarial robustness, false positive rates, class imbalance, and generalization across different IoT deployments represent persistent challenges that limit the practical effectiveness of current approaches [168][169].

The implications of these findings vary across different stakeholder communities. For security researchers in academia and industry, the survey highlights several actionable directions. Researchers should prioritize hybrid architectures that combine spatial and temporal feature extraction for traffic-based detection, with particular attention to CNN-LSTM and attention-based models that have shown superior performance across multiple studies [134][137]. They should evaluate models on multiple realistic datasets beyond single benchmark sets to assess generalization capability, and report comprehensive performance metrics including not only accuracy but also precision, recall, false positive rates, and computational costs [153][154]. Researchers should invest in adversarial robustness evaluation and defense mechanisms including adversarial training, defensive distillation, and certified defenses to ensure models remain effective against adaptive attackers [168]. They should develop explainable AI techniques adapted to security contexts that can provide interpretable explanations for classification decisions, supporting incident analysis and building trust with security analysts [171]. Researchers should explore privacy-preserving learning approaches including federated learning and differential privacy that enable collaborative model training without centralizing sensitive data [172].

For IoT system developers and security engineers responsible for implementing defensive mechanisms, the survey provides practical guidance. Developers should carefully analyze the trade-offs between detection accuracy and computational cost when selecting deep learning approaches, considering whether sophisticated models can be deployed on available hardware or whether simpler models or edge-cloud architectures are more appropriate [160]. They should leverage model compression techniques including quantization, pruning, and knowledge distillation to make sophisticated models feasible on resource-constrained devices, accepting modest accuracy degradation in exchange for substantial efficiency improvements [163]. Developers should implement comprehensive monitoring and validation of deployed models to detect performance degradation over time due to concept drift, adversarial attacks, or changes in traffic patterns [169]. They should design human-in-the-loop workflows for high-stakes security decisions where false positives or false negatives have significant consequences, using deep learning to assist rather than replace human judgment [166]. Developers should plan for regular model updates and retraining to adapt to evolving threats, incorporating mechanisms for secure model distribution and validation [106].

For IoT device manufacturers and platform providers, the survey illuminates strategic considerations. Manufacturers should consider integrating hardware acceleration such as neural processing units or specialized security processors that can enable sophisticated deep learning inference on edge devices [162]. They should develop standardized interfaces and abstractions that facilitate deployment of security models, including support for common deep learning frameworks, efficient model loading and execution, and tools for performance profiling and optimization [164]. Manufacturers should engage with the research community and standards bodies to contribute to the development of benchmark datasets, evaluation methodologies, and interoperability standards that can accelerate adoption of deep learning security solutions [150]. They should consider the full life-cycle of security models including initial deployment, monitoring, updating, and decommissioning, designing systems that support secure and efficient model management [106].

For policymakers, standards bodies, and industry stakeholders, the survey provides perspective on technology maturity and governance needs. Stakeholders should support the development of standardized benchmark datasets, evaluation protocols, and performance metrics that enable fair comparison of approaches and facilitate technology assessment, potentially through collaborative efforts involving academia, industry, and government [150]. They should encourage research into interpretable and explainable deep learning for security to support regulatory oversight, compliance verification, and incident investigation, particularly for applications in critical infrastructure and sensitive domains [171]. Stakeholders should promote privacy-preserving machine learning techniques and establish guidelines for data collection, use, and retention in IoT security applications, balancing security benefits against privacy rights [172]. They should assess the security implications of deep learning systems themselves including adversarial vulnerabilities and potential for misuse, developing standards and best practices for secure AI system design [168]. Stakeholders should facilitate technology transfer from research to commercial deployment through funding mechanisms, public-private partnerships, and support for pilot projects that demonstrate practical benefits in operational environments [103].

This survey has several important limitations that should be considered. As a secondary synthesis of published literature, the conclusions depend on the scope, quality, and reporting practices of original studies which vary considerably across the corpus. Publication bias likely results in overrepresentation of positive results and underrepresentation of negative findings and limitations. The temporal restriction to 2020-2025 provides focus on recent advances but excludes potentially relevant foundational work from earlier periods. The search strategy may have missed relevant papers in specialized venues, non-English publications, or very recent work not yet indexed. The heterogeneity of experimental conditions, datasets, and metrics limits quantitative comparison and meta-analysis. The rapid evolution of both threats and techniques means some findings may be superseded even during survey preparation.

Future research directions that emerge from this survey span technical, methodological, and systemic dimensions. From a technical perspective, developing adversarially robust deep learning architectures through adversarial training, certified defenses, and robustness verification techniques represents a critical priority given the security-critical nature of the application domain [168]. Advancing explainable AI techniques that provide interpretable explanations for security classifications including attention visualization, prototype-based explanations, and counterfactual analysis would support incident response and build trust with security practitioners [171]. Exploring federated learning and split learning architectures optimized for IoT constraints would enable privacy-preserving collaborative training while managing communication overhead and heterogeneity across devices [172]. Developing ultra-lightweight architectures and neural architecture search methods specifically tailored to IoT security tasks and resource constraints would improve deployability on edge devices [163]. Investigating continual learning and online adaptation techniques that enable models to evolve with changing threats while avoiding catastrophic forgetting would improve long-term effectiveness [106].

From a methodological perspective, establishing standardized evaluation frameworks including reference datasets spanning diverse IoT domains and attack types, agreed-upon performance metrics that capture accuracy, efficiency, robustness, and operational considerations, and protocols for fair comparison that account for different hardware platforms and deployment contexts would significantly benefit the field [150]. Conducting long-term field studies that evaluate deep learning security systems in operational environments over extended periods would provide essential evidence of practical viability and reveal challenges not apparent in laboratory settings [160]. Developing comprehensive threat models and adversarial evaluation methodologies that assess model robustness against adaptive attackers with various capabilities and objectives would improve security assurance [168]. Creating synthetic data generation and augmentation techniques that address class imbalance and improve coverage of rare attack types while maintaining realism would enhance model training [167].

From a systemic perspective, advancing privacy-preserving and secure machine learning techniques including federated learning frameworks optimized for IoT heterogeneity and constraints, differential privacy mechanisms that protect sensitive data while enabling useful inference, and secure multiparty computation approaches for collaborative learning would address critical adoption barriers [172]. Improving integration with existing security infrastructure including security information and event management systems, threat intelligence platforms, and incident response workflows would facilitate practical deployment [101]. Developing economic and risk models that quantify the costs and benefits of deep learning security solutions compared to alternatives would inform investment and deployment decisions [103]. Addressing ethical and societal implications including fairness and bias in security classifications, transparency and accountability in automated security decisions, and governance frameworks for AI-based security systems would ensure responsible development and deployment [103].

In conclusion, deep learning has emerged as a powerful paradigm for classification of security issues in IoT devices, demonstrating superior detection accuracy compared to traditional approaches while offering the potential for adaptive learning and generalization to novel threats. The field has progressed from early proof-of-concept studies to sophisticated architectures evaluated on realistic datasets, with growing attention to practical deployment challenges. Empirical evidence demonstrates that deep learning classifiers can achieve excellent performance on benchmark datasets, with hybrid architectures and careful model design yielding accuracies often exceeding 98 percent. However, significant challenges remain including computational constraints for edge deployment, adversarial vulnerabilities, false positive rates in operational environments, and generalization across diverse IoT contexts. Realizing the full potential of deep learning for IoT security will require continued innovation in architectures and training techniques, rigorous evaluation under realistic and adversarial conditions, development of supporting infrastructure and standards, and careful attention to privacy, interpretability, and practical deployment considerations. By addressing the technical challenges and research gaps identified in this survey while remaining mindful of operational constraints and broader implications, the research community can advance toward a future where intelligent, adaptive, and robust security systems protect the billions of IoT devices that increasingly underpin critical infrastructure, essential services, and daily life.

7. REFERENCES

Aldhaheri, A., Alwahedi, F., Ferrag, M. A., & Battah, A. A. (2023). Deep learning for cyber threat detection in IoT networks: A review. Internet of Things and Cyber-Physical Systems, 3, 61-77. https://doi.org/10.1016/j.iotcps.2023.09.003

Al-Garadi, M. A., Mohamed, A., Al-Ali, A., Du, X., Ali, I., & Guizani, M. (2020). A survey of machine and deep learning methods for Internet of Things (IoT) security. IEEE Communications Surveys and Tutorials, 22(3), 1646-1685. https://doi.org/10.1109/COMST.2020.2988293

Alkhudaydi, O. A., Krichen, M., & Alghamdi, A. (2023). A deep learning methodology for predicting cybersecurity attacks on the Internet of Things. Information, 14(10), 550. https://doi.org/10.3390/info14100550

Ghumro, A., Memon, A. K., Memon, I., & Simming, I. A. (2020). A review of mitigation of attacks in IoT using deep learning models. International Journal of Computer Science and Network Security, 20(6), 98-110.

Kalra, A., & Kumar, M. (2024). Classification of deep learning methods in intrusion detection for IoT devices. In 2024 International Conference on Data Science and Network Security (ICDSNS) (pp. 1-6). IEEE. https://doi.org/10.1109/icdsns62112.2024.10691071

Khan, A. R., Kashif, M., Jhaveri, R. H., Raut, R. G., Saba, T., & Bahaj, S. A. O. (2022). Deep learning for intrusion detection and security of Internet of Things (IoT): Current analysis, challenges, and possible solutions. Security and Communication Networks, 2022, Article 4016073. https://doi.org/10.1155/2022/4016073

Kornaros, G. (2022). Hardware-assisted machine learning in resource-constrained IoT environments for security: Review and future prospective. IEEE Access, 10, 58603-58622. https://doi.org/10.1109/ACCESS.2022.3179047

Tsimenidis, S., Lagkas, T., & Rantos, K. (2022). Deep learning in IoT intrusion detection. Journal of Network and Systems Management, 30(1), 1-40. https://doi.org/10.1007/S10922-021-09621-9