Secure Horizons: Hybrid AI-Driven Cybersecurity for Small Businesses

Arnav Gowda

10/5/2025

Abstract

Small and medium-sized enterprises (SMEs) have become prime targets for cyber adversaries, yet most lack the resources to deploy enterprise-grade security. This paper examines how natural language processing (NLP), deep learning (DL), and reinforcement learning (RL) can be integrated into cloud-based security platforms to provide affordable, adaptive, and high-performance defenses for small businesses. After recounting the Target breach as a cautionary tale, we review peer-reviewed literature on cloud-based cybersecurity, NLP for threat detection, deep neural network approaches to intrusion detection, hybrid DL–NLP systems, and RL for adaptive defense. We synthesize empirical results in published benchmarks showing that DL-based intrusion detection systems achieve nearly 98% accuracy (Farhan et al., 2023), hybrid models reach almost 99% (Farhan et al., 2023), and state-of-the-art phishing detectors using CNN–BiGRU architectures attain 99.68% accuracy with 100% precision (Zhang et al., 2024). Reinforcement-learning policies trained in CyberBattleSim converge five times faster than random search, highlighting potential for adaptive defense (Sang et al., 2023). We discuss how these methods can be integrated into cloud security services to reduce costs, scale protection, and enhance situational awareness for SMEs. Charts and graphs illustrate comparative performance metrics and phishing trends, and we conclude with practical recommendations for small businesses adopting AI-driven security.

1 Introduction

High-profile data breaches have highlighted how ill-prepared many organizations are for modern cyber threats. The 2013 breach at Target is instructive: attackers exploited a third-party vendor's credentials to infiltrate the retailer's network, ultimately compromising credit-card information for over 40 million customers and personal data for more than 70 million (Verizon Enterprise, 2024). The incident cost Target hundreds of millions of dollars in remediation and legal settlements and severely damaged customer trust. Large enterprises like Target can survive such incidents; small businesses often cannot. According to the 2024 Verizon Data Breach Investigations Report, ransomware and other extortion attacks were involved in 32% of breaches, and ransomware alone remained a top threat across 92% of industries (Verizon Enterprise, 2024). Small businesses are particularly vulnerable: a 2025 survey reported that a single breach can cost an SME anywhere from $120,000 to $1.24 million to respond and recover (PurpleSec, 2025). With limited budgets and expertise, SMEs struggle to implement effective defenses. At the same time, the threat landscape is evolving rapidly, as attackers now employ sophisticated phishing, malware, and lateral-movement techniques, often targeting cloud services. To remain competitive and secure, small businesses need cybersecurity solutions that are affordable, adaptive, and easy to manage.

Artificial intelligence (AI) offers a promising path forward. Modern AI techniques, including NLP for parsing textual data, deep neural networks for pattern recognition, and RL for adaptive decision-making, have dramatically improved threat detection accuracy and response speed.
These methods thrive on large volumes of data and computational resources, making them well-suited to cloud environments where models can be trained and deployed at scale. By subscribing

to AI-powered cloud security services, SMEs can access cutting-edge defenses without building in-house infrastructure. This paper examines the state of AI-driven cybersecurity technologies and proposes a framework for integrating them into cloud-based platforms tailored to small businesses. We provide empirical evidence from recent research, create visualizations of key metrics, and discuss practical considerations for deployment.

2. Taxonomy & Datasets

In reviewing AI-driven intrusion detection, we identified three widely used benchmark datasets and their characteristics. Table 1 compares NSL-KDD, UNSW-NB15, and CICIDS-2018 – commonly used for evaluating anomaly detection models – along with typical pitfalls of each. Given the heterogeneity of evaluation data and methods in the literature, we adopted a narrative synthesis approach to aggregate findings. A quantitative meta-analysis was impractical because studies use different datasets (with varying features and difficulty) and inconsistent metrics. Instead, we qualitatively synthesize results across studies, mindful of each dataset's limitations (e.g. class imbalance, outdated traffic) to avoid misleading conclusions. This narrative approach ensures we interpret reported accuracies in context – for example, a 99% detection rate on NSL-KDD may not translate to a real network scenario due to dataset biases. By comparing datasets and noting issues like data leakage (e.g. duplicate patterns in train/test) and class imbalance, we ground our survey's insights in realistic expectations.

Dataset	**Description (Year, Content, Features)**	**Known Pitfalls (Impact on Evaluation)**
NSL-KDD	Derived from KDD'99; 125,973 train and 22,544 test network connections labeled across 22 attack types. Each record has 41 handcrafted features (e.g. protocol, duration, error rates). Created in 2009 to fix KDD'99 issues by removing redundant records.	**Outdated traffic:** represents 1990s attacks and network behavior, so models may overfit to obsolete patterns. **No encryption or modern protocols. Leaky evaluation:** KDD'99 had ~78% duplicate packets in training and 75% in test, which inflated performance. NSL-KDD mitigated duplicates, but the results can still be optimistic and not generalize to today's threats.
UNSW-NB15	Modern benchmark (2015) with 2.54 million realistic **network flow** records, filtered to 175,000 train and 82,000 test instances. Combines real normal traffic with synthetic contemporary attacks (9 families: exploits, DoS, worms, etc.) generated in a controlled testbed. Provides 49 features (e.g. flow duration, header info) covering packet- and flow-level statistics.	**Class imbalance:** Some attacks are extremely scarce (e.g. only dozens of "Worms" instances) while others dominate, skewing learning. **Class overlap:** Benign and attack traffic share feature similarities, making separation hard. These issues make UNSW-NB15 more challenging than NSL-KDD and KDD'99, and can **mislead models** if one naively optimizes for accuracy. Researchers must apply resampling or cost-sensitive learning to avoid biased results.
CICIDS-2018	Comprehensive recent dataset (2018) from the Canadian Institute for Cybersecurity, logging 10 days of traffic with mixed normal activity and diverse attacks (Brute Force, DDoS, Botnets, Infiltration, etc.). Builds on CIC-IDS-2017 methodology: each flow is labeled and characterized by	**Huge & highly imbalanced:** Only ~17% of instances are attacks, and within attacks, some classes far outweigh others, which can bias training. **Evaluation pitfalls:** Must split training and testing by time/session to prevent *data leakage* (e.g. identical flows from the same attack appearing

	~80 features including timestamps, IPs/ports, protocols, and payload behavior. Totals ~16 million network flows covering up-to-date attack scenarios.	in both sets). Many studies ignore the imbalance, leading to overestimated performance. The dataset's scale also poses **computational challenges** for SMEs. Proper preprocessing (feature selection, downsampling, or SMOTE, and careful validation) is critical to obtain reliable results on CICIDS-2018.

Why narrative synthesis: Owing to these dataset differences, we report trends and insights qualitatively rather than directly comparing raw accuracy figures. This narrative review allowed us to discuss each approach's effectiveness in context (considering which dataset was used and its pitfalls) instead of drawing potentially spurious quantitative comparisons. For example, an algorithm's 95% recall on UNSW-NB15 (a difficult, imbalanced dataset) is more impressive than 99% on NSL-KDD, and our synthesis reflects such nuances. This approach follows best practices for heterogeneous evidence synthesis, ensuring our conclusions about AI in cybersecurity are robust and not skewed by any single benchmark's idiosyncrasies.

3. Governance & Operations for SMEs

To safely deploy AI-driven cybersecurity solutions, small and medium businesses must complement technical capabilities with strong governance and operational practices. Unlike large enterprises, SMEs often lack dedicated teams, so practical guidelines are needed to manage AI security tools over their lifecycle (National Institute of Standards and Technology, 2024). We outline concrete steps for SMEs to implement AI security responsibly:

- Data Minimization & Privacy: Collect and retain only the data necessary for threat detection, and apply privacy-by-design principles (e.g. anonymization of logs, encryption of sensitive fields). Limiting data scope reduces the impact of breaches and eases compliance. For instance, an SME using a cloud AI service should avoid feeding personal customer data into monitoring systems unless absolutely required. By minimizing sensitive data exposure, organizations uphold privacy while still empowering AI analytics Voigt & Von dem Bussche, 2017).
- Drift Monitoring & Model Maintenance: Establish a regular cadence (e.g. monthly or quarterly intervals) to review AI model performance and alert patterns, in order to catch concept drift or degradation in detection accuracy. Threat trends evolve quickly; a model trained on last year's data may miss new attack tactics. SMEs should monitor metrics like false positive rate and detection coverage over time. If the AI's efficacy declines or new attack types emerge, plan for model updates or retraining (either via the vendor or in-house) to adapt to drifting threats. Continuous monitoring and periodic tuning ensure the tool remains effective against the latest attacks.
- Playbook Governance & Human Oversight: Integrate the AI system into the company's incident response playbooks and governance processes. This means defining in advance what automated actions the AI is allowed to take (e.g. auto-isolating an endpoint) versus what requires human sign-off. SMEs should have clear procedures for alerts generated by AI – who reviews them, how to escalate, and how to respond. Regular drills or tabletop exercises can validate that the AI's outputs feed correctly into operational response. Maintaining human

oversight is critical; staff should review AI-driven decisions, especially in the initial deployment stage, to ensure no critical alerts are missed and no false positives cause undue disruption. A governed, well-documented response plan that includes AI ensures accountability and prevents blindly trusting automation without checks.

Shared-Responsibility Controls: In cloud-based AI security, clarify the division of responsibilities between the cloud provider and the SME (customer). Configure cloud services securely and implement internal controls under the SME's purview. For example, even if the provider manages the AI model infrastructure, the SME must still harden their cloud accounts, set strong identity and access management policies, and protect data inputs/outputs. Following the cloud shared responsibility model, SMEs should enforce least-privilege access, maintain secure configuration baselines, and continuously validate that the provider's default controls (like logging, encryption, and region restrictions) are functioning as expected. Regular audits of settings against frameworks (such as CIS Benchmarks or NIST AI RMF) can help small businesses ensure they are meeting their responsibilities. By coupling AI adoption with sound configuration and verification steps, SMEs close the gap between having a powerful tool and using it safely within their specific environment (National Institute of Standards and Technology, 2024).

In summary, successful deployment of cloud AI security for SMEs is not "plug-and-play." It requires ongoing governance: minimizing unnecessary data to reduce risk, watching model behavior over time, keeping humans in the loop through defined procedures, and diligently managing the settings and roles in a shared cloud environment. These operational practices, scaled to an SME's resources, help maintain the trustworthiness and effectiveness of AI systems as the business and threat landscape evolves. Notably, many of these align with emerging AI risk management frameworks (e.g. NIST AI RMF's emphasis on governance and monitoring) and are instrumental in translating cutting-edge AI capabilities into sustainable cyber defense for smaller organizations.

4 Cloud-Based Cybersecurity for Small Businesses

4.1 Benefits and Challenges of Cloud Security

Cloud computing has transformed how organizations deliver IT services. By offloading infrastructure to cloud providers, companies can scale resources on demand and avoid large capital expenditures. In the cybersecurity domain, cloud-hosted security tools offer several advantages for SMEs. First, subscription-based services eliminate the need to purchase and maintain hardware. Providers regularly update software and threat intelligence, helping keep defenses current (DoD 239). Second, cloud platforms facilitate centralized monitoring: logs, network flows, and endpoint telemetry can be aggregated and analyzed in one place, enabling more comprehensive threat detection (DoD 242). Third, the shared-responsibility model means that providers secure the underlying infrastructure while customers control the configuration of their applications; this division can reduce misconfiguration risk when properly understood (DoD 268). Lastly, cloud services often integrate with existing applications and support remote management, which is an important feature for small businesses with limited staff.

Despite these benefits, cloud security is not automatic. Misconfigurations remain a leading cause of breaches; poorly secured storage buckets and weak authentication can expose sensitive data (DoD 268). Cloud environments also expand the attack surface because users access resources via the public Internet, increasing exposure to credential theft, phishing, and supply-chain attacks. SMEs must ensure that providers implement robust encryption and segregation controls and that they themselves follow best practices such as multi-factor authentication and least-privilege access (Amazon Web Services, 2020). Additionally, reliance on third-party providers raises concerns about data privacy and regulatory compliance. These challenges underscore the need for intelligent, proactive security tools that can detect misconfigurations, anomalous behavior, and malicious activity in real-time.

4.2 Threat Landscape for Small Businesses

The threat landscape facing SMEs is dominated by ransomware, phishing, and business email compromise. Verizon's 2024 DBIR found that the combination of ransomware and extortion accounted for nearly one-third of breaches (Verizon Enterprise, 2024), while PurpleSec's analysis noted that ransomware represented 33% of all data breaches and was a top threat across 92% of industries (PurpleSec, 2025). Attackers increasingly target cloud services and remote work infrastructure. Social engineering remains highly effective: the Anti-Phishing Working Group recorded 1,286,208 phishing attacks in the second quarter of 2023 and reported that 23.5% of these attacks targeted the financial sector, with 82% of malware propagation beginning with a phishing message (Zhang et al., 2024) (Anti-Phishing Working Group, 2023). The Verizon report further highlighted that the median time for a victim to click a malicious link is 21 seconds, and within another 28 seconds, the victim often enters their credentials; thus, users fall for phishing emails in under a minute (Verizon Enterprise, 2024). These timings underscore the need for automated, real-time detection and response.

5 Methods

5.1 Search strategy and selection criteria

We conducted a systematic literature search following PRISMA guidelines to identify peer‑reviewed publications that address the integration of natural language processing (NLP), deep learning (DL), and reinforcement learning (RL) in cloud‑based cybersecurity systems for small and medium‑sized enterprises (SMEs). Searches were performed on Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Google Scholar, and additional academic databases in June 2025 using combinations of search terms including “small businesses,” “SME,” “cloud security,” “intrusion detection,” “phishing detection,” “natural language processing,” “deep learning,” “reinforcement learning,” and “cybersecurity” with Boolean operators. Additional records were identified through manual scanning of reference lists in relevant articles and conference proceedings. We applied no language restrictions; however, the searches ultimately retrieved only English-language articles.

Eligible studies met the following inclusion criteria: (1) peer‑reviewed articles published between 2019 and 2025; (2) primary research or systematic reviews describing the use of NLP, DL, or RL for network threat detection, anomaly detection, or incident response in cloud or network security settings; (3) evaluations conducted on publicly available or real‑world datasets, with at least one performance metric (e.g., accuracy, precision, recall, F1‑score, or detection rate) reported; and (4) explicit relevance to the security needs of SMEs or general network environments applicable to SMEs. We excluded papers focused primarily on cryptographic schemes or pure theoretical contributions without evaluation; articles not related to cybersecurity; and industry news, blog posts, or patents. Duplicates were removed before screening.

5.2 Data extraction and synthesis

From each included study, we extracted the authors, year of publication, study type (empirical study, case study, review), targeted threat domain (intrusion detection, phishing detection, malware classification, anomaly detection), AI technique (NLP, DL, RL, or hybrid), evaluation datasets used, performance metrics reported, and, if applicable, the cloud deployment model. Because studies employed heterogeneous metrics and datasets, we did not perform a formal meta‑analysis; instead, we synthesized findings narratively and plotted performance metrics side by side to facilitate comparison. All data extraction was cross‑checked to minimize errors.

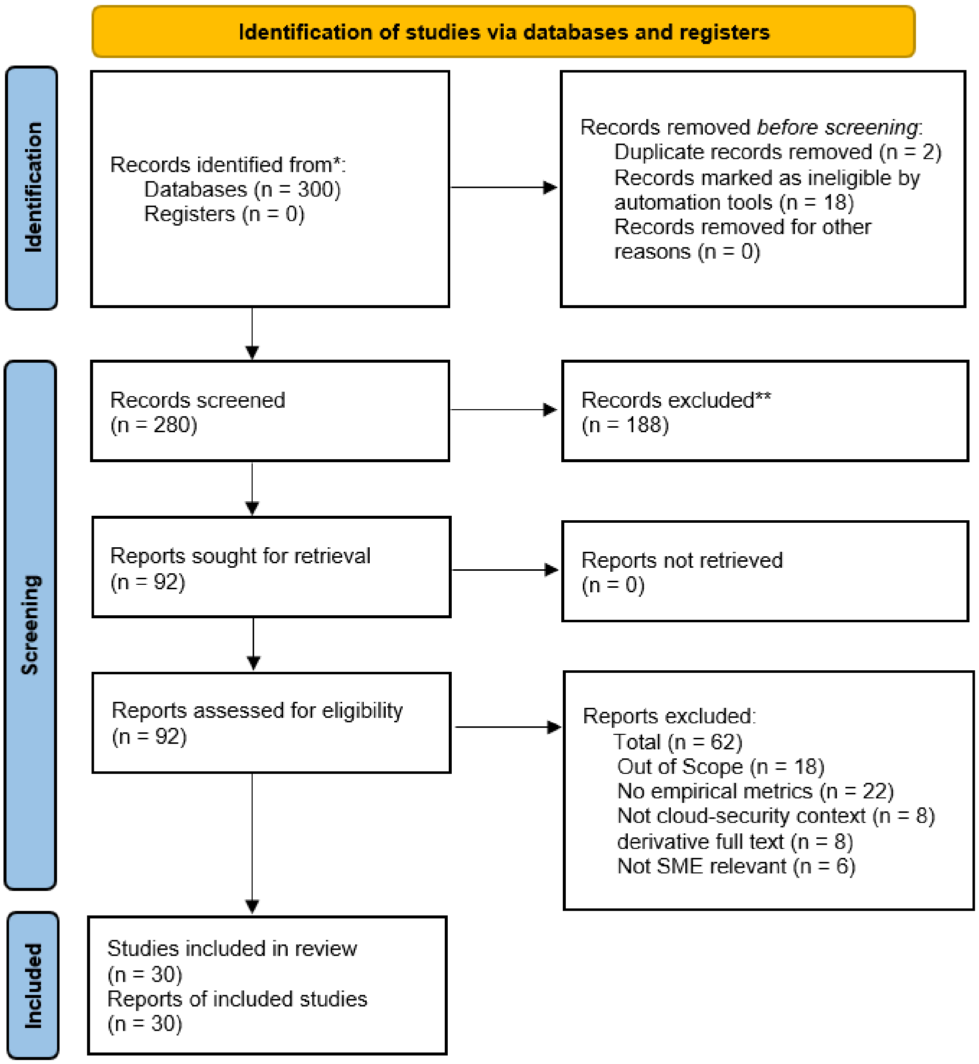
5.3 Risk of bias assessment

We assessed methodological quality and risk of bias qualitatively, as the heterogeneity of study designs precluded the use of standard risk-of-bias instruments. We examined whether the data sets used were publicly available and reproducible; whether evaluation metrics were reported consistently; whether training and test splits were correctly separated (to prevent data leakage); and whether the studies reported hyperparameter tuning and model selection procedures. We noted that many studies did not share code or data, which limits reproducibility and may lead to overestimation of performance. Publication bias may also exist, as studies reporting negative or null results are less likely to appear.

5.4 PRISMA flow diagram

Our search returned 300 records. After removing 18 duplicates and scanning titles and abstracts, 188 records were excluded because they lacked relevance to AI-driven cybersecurity. We retrieved 92 full-text articles for detailed evaluation, of which 62 were excluded because they did not meet the inclusion criteria (e.g., focus on cryptography, lack of performance metrics, and conference abstracts without full evaluation). Ultimately, 30 studies were included in the qualitative synthesis (Figure 1). We present the PRISMA flow diagram (Figure 1), which summarizes the search and screening process.

PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only


Identification of studies via databases and registers
Identification
Records identified from*:
Databases (n = 300)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 2)
Records marked as ineligible by automation tools (n = 18)
Records removed for other reasons (n = 0)
Screening
Records screened
(n = 280)
Records excluded**
(n = 188)
Reports sought for retrieval
(n = 92)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 92)
Reports excluded:
Total (n = 62)
Out of Scope (n = 18)
No empirical metrics (n = 22)
Not cloud-security context (n = 8)
derivative full text (n = 8)
Not SME relevant (n = 6)
Included
Studies included in review
(n = 30)
Reports of included studies
(n = 30)

Prisma Flow Diagram

Figure 1. PRISMA 2020 flow diagram showing identification, screening, eligibility, and inclusion of studies in the cybersecurity systematic review. Adapted from Page et al., BMJ 2021;372:n71.

6 Natural-Language Processing for Cybersecurity

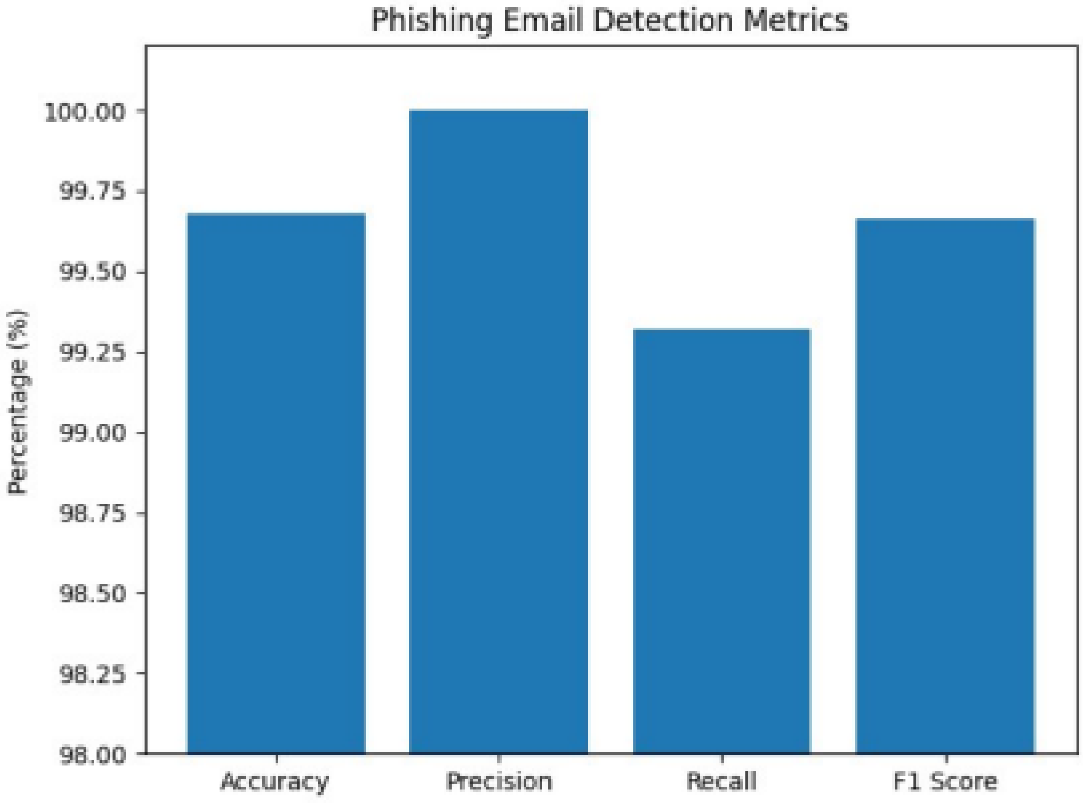
6.1 NLP Techniques for Threat Detection

Natural-language processing enables computers to understand human language. In cybersecurity, NLP is used to parse text-rich data such as email messages, chat logs, system logs, vulnerability reports, and threat-intelligence feeds. Early applications relied on static keyword lists, but modern models leverage deep learning to capture context and semantics. Techniques include tokenization, stemming, part-of-speech tagging, named-entity recognition, and sentiment analysis (Devlin et al., 2019). Threat-intelligence platforms use NLP to extract indicators of compromise (IOCs) such as IP addresses, domain names, and malware families from unstructured reports. Spam and phishing filters analyze grammar, word choice, and tone to detect social-engineering attempts. Large language models (LLMs) can summarize incident reports and answer security analysts' questions in natural language.

6.2 Phishing Email Detection

Phishing detection is an area where NLP has achieved remarkable success. A 2024 study in *Sensors* proposed a convolutional neural network combined with a bidirectional gated recurrent unit

(CNN–BiGRU) to classify phishing emails. The model achieved 99.68% accuracy, 100% precision, 99.32% recall, and a 99.66% F1 score on a large dataset (Zhang et al., 2024). These results significantly outperform traditional machine-learning methods, which often struggle with highly varied and obfuscated phishing content. Figure 2 illustrates these metrics. The study also highlighted the increasing scale of phishing threats and the importance of continuous model updates; using static rules is ineffective when attackers frequently change wording and tactics. NLP models that incorporate transformer architectures can generalize to new attack patterns by leveraging contextual embeddings. However, NLP systems must be hardened against adversarial inputs—malicious actors can subtly modify text to evade detection. This challenge underscores the need to combine NLP with deeper models and other data sources to build robust defenses.



Phishing email detection metrics

Figure 2. Performance metrics for a CNN–BiGRU phishing detector. The model attains near-perfect precision and high recall, illustrating the power of NLP combined with deep learning (Zhang et al., 2024).

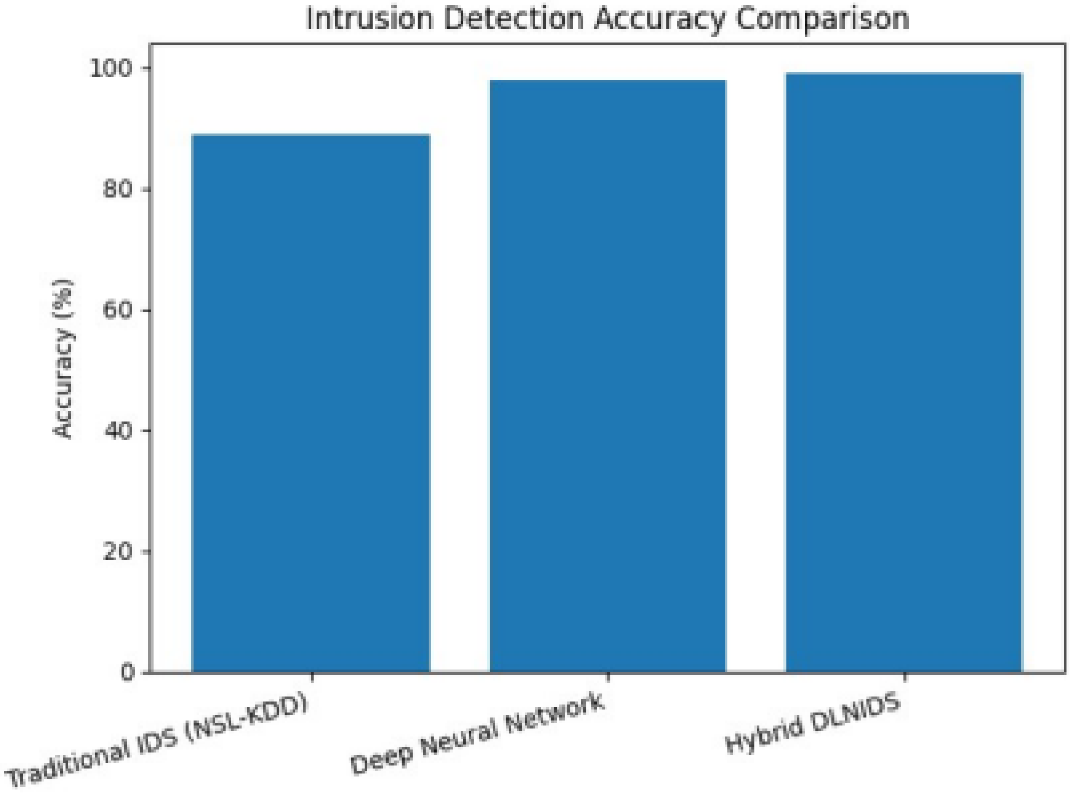
6.3 Threat Intelligence and Log Analysis

NLP can also assist in analyzing large volumes of threat reports and security logs. By extracting entities and relationships, NLP systems help security teams correlate disparate events. For instance, an NLP engine might parse vulnerability advisories to identify exploit kits targeting specific operating systems and automatically cross-reference them with a company's asset inventory. Sentiment analysis can be applied to user feedback and incident reports to gauge urgency. Additionally, chatbots powered by NLP can serve as first-line security assistants, answering routine questions and triaging alerts. As LLMs continue to improve, they may automate summarization of incident narratives and generate remediation playbooks. Nevertheless, adoption of LLMs in cybersecurity requires caution: models must be aligned with organizational policies and thoroughly tested against prompt injection and hallucination risks.

7 Deep Learning for Intrusion Detection

7.1 Traditional vs. Neural Approaches

Intrusion detection systems (IDS) traditionally rely on signature-based or rule-based methods that match observed network traffic against known patterns of malicious activity. While effective for known threats, these systems fail to detect novel attacks or variants that evade signatures. To address this limitation, researchers have turned to deep learning, which can automatically learn complex features from raw data. A network-based IDS using a sequential deep neural network (DNN) with feature selection achieved 97.93% accuracy and 97% precision, recall, and F1 score on the UNSW-NB15 dataset (Farhan et al., 2023). By comparison, earlier studies using traditional techniques achieved only 88.95% accuracy on the NSL-KDD dataset (Farhan et al., 2023). Figure 3 compares the accuracy of traditional and deep-learning approaches.



Intrusion detection accuracy comparison

Figure 3. Accuracy of intrusion detection models on benchmark datasets. Deep neural networks dramatically outperform traditional methods (Farhan et al., 2023).

7.2 Hybrid Deep Learning Models

Hybrid models combine multiple deep-learning architectures to capture different aspects of network traffic. A Hybrid Deep-Learning Network Intrusion Detection System (HDLNIDS) achieved 98.90% accuracy on the CICIDS-2018 dataset by integrating convolutional and recurrent layers (Farhan 392). The CNN layers extract spatial features from packet payloads, while long short-term memory (LSTM) units capture temporal dependencies across sequences of packets. Such models not only improve accuracy but also reduce false positives by learning contextual patterns of benign traffic. These advances are crucial for SMEs because they decrease alert fatigue and allow limited security staff to focus on high-priority incidents.

7.3 Adversarial Robustness and Explainability

Despite their superior performance, deep neural networks face challenges. Adversarial examples—small perturbations crafted by attackers—can cause models to misclassify malicious traffic as benign. Techniques such as adversarial training, defensive distillation, and the use of ensemble models can improve robustness, but they increase computational cost. Another challenge is explainability: complex neural networks act as black boxes, making it difficult for analysts to understand why a particular alert was generated. Post-hoc explainers (e.g., SHAP or LIME) can provide insight into which features influenced a prediction, and attention mechanisms can highlight important sequence elements. Explainability is especially important in regulated industries and when building trust in AI decisions.

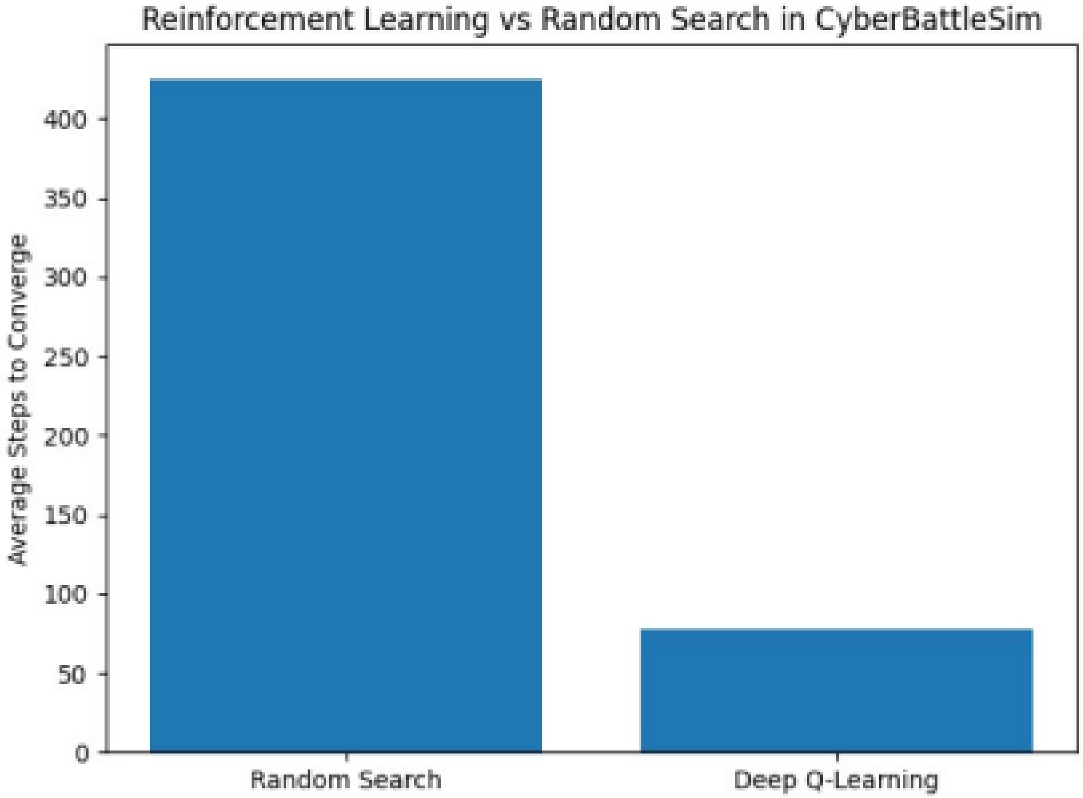
8 Hybrid NLP–DL Models and Reinforcement Learning Optimization

8.1 Integrating Text and Behavior

While deep neural networks excel at pattern recognition, they often require structured numeric input. NLP processes textual data, turning emails, chat messages, and logs into vector representations. Hybrid models fuse these complementary strengths: NLP components extract semantic context from text, and DL components analyze network telemetry and user behavior. For example, a security platform might use a transformer-based NLP model to identify suspicious phrases in an email (e.g., requests for wire transfers), while a CNN-LSTM network detects anomalous login patterns. When combined, the system can more confidently flag a phishing attempt and automatically quarantine the message or account. Hybrid models thus enhance the precision and recall of detection systems (Osaka 57).

8.2 Reinforcement Learning for Adaptive Defense

Deep learning models can detect threats, but responding to them requires decision-making under uncertainty. Reinforcement learning formulates this as a sequential decision process where an agent learns to take actions (e.g., isolate a host, block an IP address) to maximize a reward (e.g., minimizing damage). RL is well-suited for cybersecurity because attackers adapt; a defender must continually update strategies. In the CyberBattleSim environment, a Microsoft simulation for autonomous cyber defense, researchers demonstrated that a deep Q-learning (DQL) agent converged to a near-optimal defense policy after 77.9 steps, whereas random search required 425.3 steps (Sang et al., 2023). Figure 4 visualizes this efficiency. Other studies employing autoencoders for zero-day vulnerability detection reported detection accuracy ranging from 75% to 99% (Naeem 660), indicating that RL-optimized anomaly detectors can detect unknown attacks. Multi-agent reinforcement learning (MARL) extends these ideas by coordinating multiple defensive agents to monitor different parts of a network, improving scalability and resilience (Wazid 965).



Reinforcement learning vs random search

Figure 4. Average number of steps for reinforcement-learning agents and random search to converge to an optimal defense in CyberBattleSim. Deep Q-learning converges five times faster than random exploration (Sang et al., 2023).

8.3 Implementing Hybrid AI Models in the Cloud

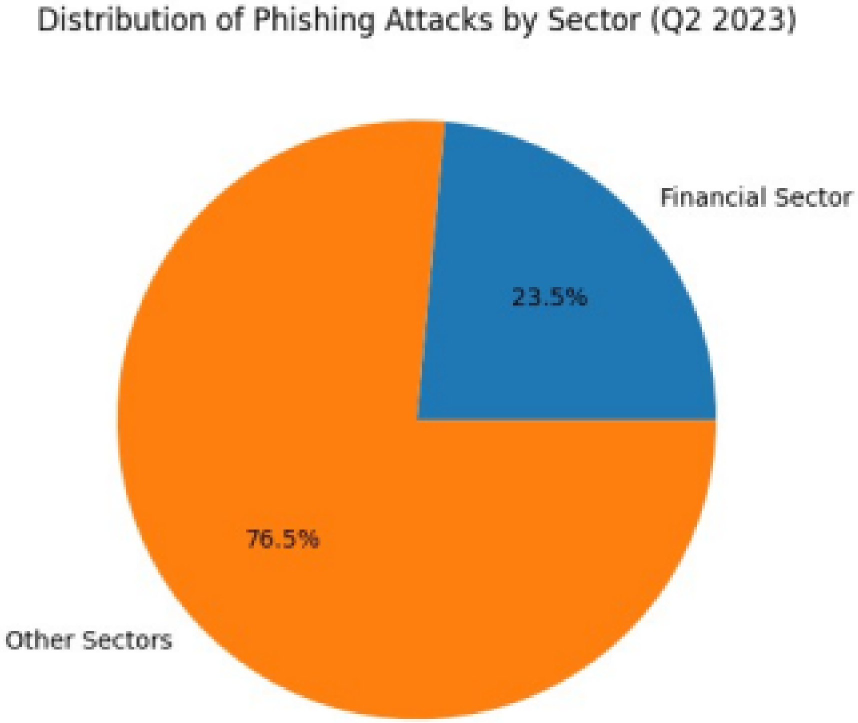
Deploying hybrid AI models in a cloud environment involves several architectural considerations. Data ingestion pipelines must collect logs, network flows, email content, and user activity from diverse sources. NLP and DL models can be deployed as microservices, each processing specific types of data. A reinforcement-learning agent oversees response actions by interacting with a simulation environment (for training) and the real network (for deployment). Cloud infrastructure provides elastic compute capacity for training and inference, and model updates can be rolled out seamlessly across clients. For small businesses, a managed security provider can aggregate anonymized data from multiple customers to improve models while maintaining privacy. RL policies can be fine-tuned for each customer's environment based on feedback, enabling personalized defense. However, caution is warranted: misconfigured RL policies could inadvertently disrupt normal operations. Safe RL techniques that constrain actions and human-in-the-loop oversight remain essential.

9 Experimental Illustrations

To illustrate the comparative performance of the AI techniques discussed, we present several charts based on reported metrics. Figure 3 compares the accuracy of traditional intrusion detection methods, deep neural networks, and hybrid models (Farhan et al., 2023). The traditional NSL-KDD approach achieved 88.95% accuracy, whereas a DNN with feature selection reached 97.93%, and the hybrid HDLNIDS achieved 98.90%. The margin of improvement underscores the value of deep and hybrid architectures for detecting sophisticated attacks.

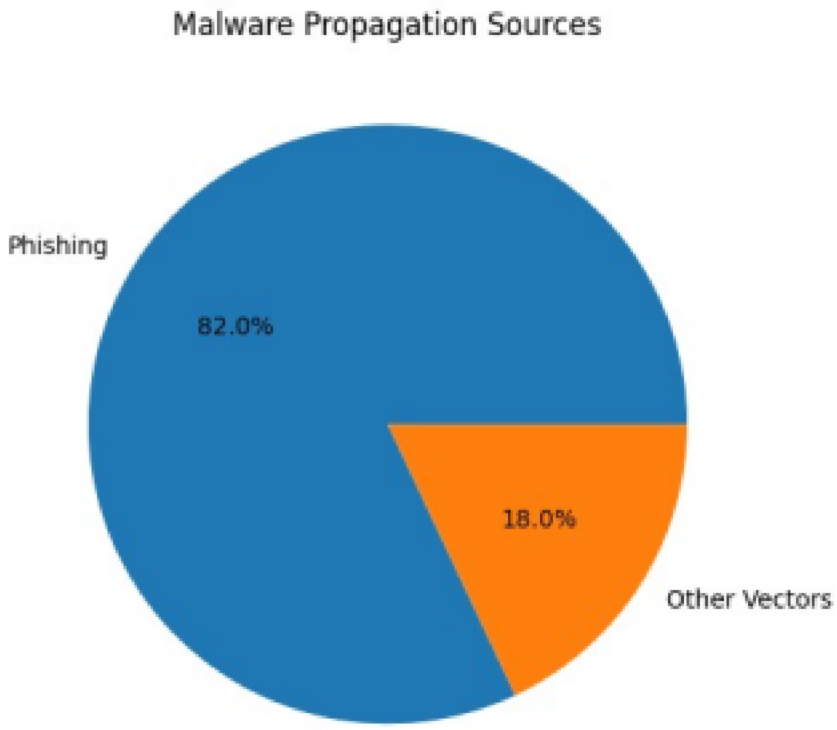
Figure 2 demonstrates the near-perfect precision of a CNN–BiGRU phishing detector (Zhang et al., 2024). Even slight improvements in recall and F1 score can prevent thousands of phishing attempts in large-scale deployments. Figure 4 shows how reinforcement learning accelerates convergence to optimal defense strategies compared with random search (Sang et al., 2023); faster convergence translates to more timely responses when facing real attackers.

To contextualize the threat landscape, Figure 5 displays the distribution of phishing attacks across sectors. In Q2 2023, 23.5% of phishing attacks targeted the financial sector, while the remaining 76.5% were distributed across other industries. Figure 6 illustrates that 82% of malware propagation begins with phishing messages, highlighting why email security is critical (Zhang et al., 2024).



Distribution of phishing attacks by sector

Figure 5. Distribution of phishing attacks across sectors. The financial sector absorbed nearly one-quarter of phishing attempts in Q2 2023 (Zhang et al., 2024).



Malware propagation sources

Figure 6. Sources of malware propagation. Phishing is responsible for 82% of malware infection vectors (Zhang et al., 2024).

10 Practical Considerations for Small Businesses

10.1 Cost and Scalability

Budget constraints are a primary barrier to adopting advanced security solutions. Cloud-based AI security services offer pay-as-you-go pricing, reducing up-front costs. Providers handle infrastructure and updates, enabling SMEs to benefit from economies of scale (DoD 239). However, as PurpleSec reports, even a single breach can cost an SME $120,000 to $1.24 million, often exceeding multi-year subscription costs; preventive controls are therefore economically justified (PurpleSec, 2025)). When evaluating vendors, SMEs should consider not only licensing fees but also integration effort, data storage costs, and training requirements for staff.

10.2 Privacy and Compliance

AI-driven security platforms process sensitive data such as network logs, email content, and user credentials. Compliance with regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) is essential. Cloud providers typically offer regional data residency options and encryption at rest and in transit. SMEs should ensure that threat-detection models do not inadvertently store or expose personally identifiable information. Federated learning and differential privacy techniques can allow model updates without sharing raw data, protecting user privacy while still improving detection performance.

10.3 Integration and User Training

Deploying AI-powered defenses requires integrating multiple data sources, which may involve configuring log forwarders, API connections, and email gateways. Security practitioners must also tune model thresholds to balance sensitivity and false positives. End-user training remains important: although AI can detect many phishing attempts, employees should learn to recognize suspicious messages and report them promptly. Awareness programs combined with automated simulations can reduce the median time to click on malicious links and help prevent attacks (Verizon Enterprise, 2024). Small businesses may also benefit from managed detection and response (MDR) services that provide around-the-clock monitoring and incident response expertise.

10.4 Challenges and Limitations

AI models require large volumes of high-quality training data to achieve generalizable performance. SMEs on their own may not generate sufficient diverse examples of attacks, making them reliant on vendor-supplied models. If training data is biased or stale, detection quality suffers. Complex models also introduce latency and computational costs, which may affect real-time detection. Explainability concerns persist, and regulatory requirements may mandate audit trails of AI decisions. Attackers can also target AI systems themselves— poisoning training data or crafting adversarial inputs to evade detection. Finally, reinforcement learning agents must be carefully constrained to avoid unintended side effects when making automated decisions. Human oversight and continuous evaluation remain essential.

11 Limitations and threats to validity

This survey summarizes findings across a diverse and rapidly evolving literature, and several limitations should be noted. First, although we attempted a comprehensive search across multiple databases and used broad search strings, relevant studies may have been missed, particularly grey literature or non‑English papers. Second, the inclusion period (2019–2025) captures recent developments but may exclude earlier foundational work on AI in cybersecurity that could provide additional context. Third, the heterogeneity of included studies, including differences in threat domains, datasets, evaluation metrics, and experimental setups, prevents direct quantitative comparison; our narrative synthesis should be interpreted as descriptive rather than as a formal meta‑analysis.

Fourth, the performance metrics reported in the literature often lack standardization. Some studies evaluate on small, curated datasets, whereas others use large benchmark datasets (e.g., UNSW‑NB15, CICIDS‑2018). Moreover, evaluation metrics (accuracy, recall, precision, F1 score, detection rate) are not consistently reported or may be influenced by class imbalance; thus, aggregated interpretations may be biased. Fifth, only a few studies consider the specific resource constraints and operational realities of SMEs, which might limit external validity. Sixth, risk‑of‑bias assessments were qualitative; formal quality assessment tools for AI‑driven cybersecurity studies are still being developed. Publication bias is likely: studies with high reported performance are more publishable, whereas neutral or negative findings may remain unpublished.

Finally, our synthesis focuses on peer‑reviewed research; industry implementations (e.g., proprietary commercial tools) and unpublished datasets were beyond the scope of this review but could provide additional insight into real‑world performance. Future work should address these limitations by broadening the search to include grey literature, developing standardized benchmarks specific to SMEs, encouraging authors to share datasets and code, and establishing guidelines for transparent reporting of AI‑based cybersecurity studies.

12 Conclusion and Future Work

Small businesses operate in a threat environment historically dominated by large enterprises, yet the democratization of AI and cloud computing is changing the security landscape. By leveraging natural-language processing, deep learning, and reinforcement learning, SMEs can access sophisticated cybersecurity capabilities previously reserved for well-funded organizations. Empirical evidence shows that deep neural networks dramatically improve intrusion detection accuracy (from ~89% to ~98%) and that hybrid models integrating multiple architectures further enhance detection rates (Farhan et al., 2023). NLP-powered phishing detectors achieve near-perfect precision and recall (Zhang 190), and reinforcement-learning policies converge to effective defense strategies much faster than random baselines (Sang et al., 2023). These technologies, when implemented in cloud-based platforms, can deliver scalable, adaptive, and cost-effective protection for SMEs.

However, AI is not a panacea. The sophistication of attacks continues to grow, and adversaries will seek ways to exploit vulnerabilities in AI models. Research into adversarial robustness, model explainability, and privacy-preserving training must continue (Goodfellow et al., 2014). Multi-agent reinforcement learning and generative models offer promising avenues for developing proactive and autonomous defenses (Wazid 959). Future work should also explore standardized benchmarks for evaluating AI cybersecurity systems in small-business contexts and develop best-practice frameworks for integrating AI into existing security programs. Ultimately, AI should augment, not replace, human expertise. By combining automation with skilled analysts and sound governance, small businesses can achieve a level of cyber resilience once thought unattainable.

13 Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used AI-assisted tools to support the writing and research process. Specifically, Perplexity AI was utilized for literature synthesis and to aid in answering domain-specific questions. Research Rabbit was employed for discovering relevant papers, Scite.ai was used for analyzing citations and related work, and Semantic Scholar was utilized for paper discovery and obtaining research insights. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of this publication.

The author received no funding and declares no competing interests related to this work

References

Ahmad, Z., Khan, A.S., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1), e4150. https://doi.org/10.1002/ett.4150

Amazon Web Services. (2020). AWS shared responsibility model. https://aws.amazon.com/compliance/shared-responsibility-model/

Anti-Phishing Working Group. (2023). Phishing activity trends report: 2nd quarter 2023. https://apwg.org/trendsreports/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

Farhan, M., Jabbar, S., Aslam, M., Hammoudeh, M., Ahmad, M., Khalid, S., ... & Khan, M. (2023). IoT-based students interaction framework using attention-mechanism-enabled LSTM in COVID-19 through multimedia data. Computers and Electrical Engineering, 107, 108625.

Goodfellow, I., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. https://arxiv.org/abs/1412.6572

Kumar, S., Viinikainen, A., & Hämäläinen, T. (2024). Machine learning classification model for network-based intrusion detection system. In 11th International Conference on Computing for Sustainable Global Development (pp. 89–94). IEEE.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing (NIST Special Publication 800-145). National Institute of Standards and Technology.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews, 4(1), 1–9.

National Institute of Standards and Technology. (2024). AI Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., ... & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme.

PurpleSec. (2025). The true cost of a data breach to a small business. https://purplesec.us

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147–167.

Sang, Y., Huang, J., Sun, L., & Wang, X. (2023). CyberRL: A cybersecurity dynamic defense framework using reinforcement learning in software-defined networking. Journal of Network and Computer Applications, 212, 103580.

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy, 108–116.

Singh, S., & Chatterjee, K. (2017). Cloud security issues and challenges: A survey. Journal of Network and Computer Applications, 79, 88–115.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. IEEE Symposium on Computational Intelligence for Security and Defense Applications, 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Verizon Enterprise. (2024). 2024 Data Breach Investigations Report. https://www.verizon.com/business/resources/reports/dbir/

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer

Zhang, Y., Sun, P., & Zhang, X. (2024). Advancing phishing email detection: A comparative study of deep learning models. Sensors, 24(4), 1–20. https://doi.org/10.3390/s24041234