Journal of Information Security Research

DLINE JOURNALS

Print ISSN: 0976-4143
Online ISSN: 0976-4151

JISR 2026: 17 (4)
https://doi.org/10.6025/jisr/2026/17/4/191-205

Extracting Cybersecurity Knowledge from Educational Visuals: An OCR-Based Concept Analysis and Semantic Knowledge-Graph Approach

Ricardo Rodríguez Jorge
Research Scientist, Ceit
Spain
ricardo.rodriguezjorge.mx@ieee.org

ABSTRACT

Unstructured cybersecurity educational visuals, such as training posters and dashboards, contain dense domain knowledge but remain difficult to reuse programmatically. Traditional text only extraction pipelines often discard critical contextual cues like layout, iconography, and visual hierarchy. This study investigates a lightweight, multimodal pipeline designed to convert minimal, image only educational resources into structured, machine readable semantic knowledge. Using a descriptive case study approach, we applied a sequential process comprising optical character recognition (OCR), thematic concept grouping, semantic knowledge graph construction, and low level colorimetric analysis to a single CompTIA CloudNetX (CNX-001) educational poster. The pipeline successfully extracted and categorized 14 core cybersecurity concepts across five foundational domains, including network protection, secure connectivity, and monitoring. Furthermore, it inferred 13 directed semantic relationships, transforming flat textual labels into a dynamic, interconnected ontology. Parallel image forensics quantified the visual metadata, revealing a "cool toned," dark mode aesthetic characteristic of modern security dashboards. The findings reflect that visual hierarchy and spatial layout can be reliably interpreted to map pedagogical and technical dependencies. This reproducible framework provides a scalable alternative to manual extraction, offering valuable machine-readable artefacts for curriculum analytics, semantic search, and multimodal reasoning in cybersecurity training.

Keywords: Cybersecurity Education, Optical Character Recognition (OCR), Semantic Knowledge Graph, Multimodal Pipeline, Unstructured Visual Data, Information Extraction, Ontology Construction, Image Forensics, Curriculum Analytics

Received: 19 March 2026, Revised 3 July 2026, Accepted 22 July 2026

Copyright: DLINE

1. Introduction

Unstructured data constitutes a critical resource for many enterprises. Manual processing of such data frequently introduces errors, delays, and reduced operational efficiency. Automatic data extraction addresses these limitations by delivering digital, automated solutions for document processing. By automating repetitive

and time intensive tasks, organizations can enhance productivity and scale their operations. Artificial intelligence enables rapid, automated extraction of relevant information from unstructured sources and supports the generation of analyses that facilitate informed decision making [1]. Digital text recognition has advanced rapidly, providing systems that bridge human computer communication by encoding printed or handwritten text into digital form. Surveys of unstructured data modalities including images, text, audio, and video highlight persistent challenges in representation, conversion to structured formats, and the exponential growth and heterogeneity of data volumes [2, 3]. Parallel structured literature reviews have catalogued the principal challenges, themes, and open research questions surrounding robotic process automation (RPA) [4].

2. Related studies

Recent work has demonstrated the utility of generative AI for converting unstructured business content into structured semantic knowledge. Pawe³oszek [5] applied large language models (LLMs) and prompt engineering to the 2025 Logility Market Report, automatically producing RDF triples that capture key entities and relationships in AI-driven supply chain management. The pipeline incorporated both pure text extraction and OCR based processing of an accompanying infographic, illustrating the capacity of multimodal LLMs to transform visual material into machine readable structures [Pawe³oszek]. In a complementary educational setting, Yi proposed a light weight screen recording analysis method for cyber range environments that combines video frame extraction, OCR-based command recognition, and the Levenshtein distance algorithm. The approach quantifies student performance during penetration testing exercises [6].

The cyber threat landscape itself remains highly dynamic, characterized by the continual emergence of new vulnerabilities, malware variants, and attack techniques. Adversaries routinely alter their behaviors to evade known indicators of compromise (IoCs) and tactics, techniques, and procedures (TTPs), thereby introducing previously unseen threat patterns [7]. High impact zero day exploits such as Log4Shell and Spring4Shell can appear with little warning and cause widespread disruption before patches become available [8]. These rapid developments strain the generalization capacity of traditional cyber threat intelligence (CTI) models that rely on handcrafted features or static scoring mechanisms such as CVSS [9].

Transfer learning has achieved notable success in natural language processing by leveraging pretrained language models (PLMs) BERT, RoBERTa, GPT, and their derivatives trained on large general-domain corpora. These models learn universal contextual representations that can be efficiently adapted to downstream tasks including text classification, named entity recognition (NER), and question answering [10, 11]. In contrast to end to end models trained from scratch, the pretrain finetune paradigm requires only limited parameter updates, reducing dependence on large labeled datasets while attaining state of the art performance. Within CTI, transfer learning has been applied successfully to document classification [12, 13] and threat entity recognition [14, 15, 16].

Nevertheless, learning robust features from cyber threat text remains challenging. CTI sources incident reports, threat bulletins, and social media feeds are characteristically noisy, heterogeneous, and replete with domain specific jargon, acronyms, and abbreviated references. Conventional statistical representations such as bag of words fail to capture the necessary semantic and contextual information [17].

Semantic knowledge graphs (SKGs) provide a powerful mechanism for identifying and categorizing entities within unstructured documents, thereby supporting knowledge discovery, graph construction, and information retrieval [18, 19, 29]. While general domain graphs focus on broad entity classes, domain specific graphs have become prominent in biomedicine, legal text processing, materials science, and, increasingly, CTI [21].

Cybersecurity knowledge graph technology has progressed from early rule based systems limited in their ability to handle multi source heterogeneous intelligence and to adapt to evolving threats to intelligent learning paradigms. Contemporary approaches exploit deep learning and generative models to enable dynamic threat modeling. Zhao et al. survey this evolution, emphasizing the shift toward neural architectures, pretrained language models, and hybrid knowledge graph LLM frameworks [22].

Concrete methodological advances illustrate this trajectory. Jia et al. constructed a cybersecurity knowledge base on a five element model, employing Stanford NER augmented with gazetteer features for domain specific entity recognition and ontology construction [23]. Han and Wang [24] improved multi entity relation extraction in long texts through entity masking combined with BERT embeddings. Mouiche and Saad [25] introduced the TiKG framework, which integrates SecureBERT domain adaptive pretraining with attention enhanced bidirectional LSTMs; the architecture achieves accurate advanced persistent threat (APT) entity identification and low error relation propagation under domain-specific ontological constraints, demonstrating improved accuracy on four benchmark datasets including DNRTI. Du et al. [26] performed vulnerability knowledge fusion via BLSTM CRF attribute extraction and hierarchical rule based relationship triggering. Li et al. proposed a BiLSTM graph neural network hybrid that jointly captures sequential context and performs graph based relation extraction [27]. Sangher et al. [28] showed that LSTM BERT pipelines can surface implicit exploitation cues from social platforms and darknet forums, enriching CTI collections.

More recent LLM centric frameworks further advance end to end extraction. Hu et al. developed LLM-TIKG, employing GPT powered few shot learning to generate labeled data for fine tuning a 7B Llama 2 model, thereby extracting threat intelligence entity TTP relation triples [29]. Zhang et al. introduced AttackKG+, a collaborative multi module LLM pipeline (rewriter parser identifier digester) that constructs multidimensional threat maps aligned with MITRE TTPs [30]. Huang and Xiao's CKG method processes multi source CTI articles through a segmented multi LLM agent architecture equipped with dual memory mechanisms, producing traceable entity–relationship graphs [31]. Paul et al. extended these capabilities with proactive LLM-based threat reasoning that infers attacker goals and behavioral chains [32], while Wu et al. incorporated knowledge-graph constraints into LLM outputs to enhance the credibility of extracted intelligence [33].

These developments establish a coherent progression from traditional OCR and rule based extraction, through transfer learning and neural architectures, to hybrid LLM knowledge graph systems capable of transforming both textual and visual cybersecurity educational materials into structured, actionable semantic knowledge. [34, 35]

3. Research Statement

Unstructured educational visuals such as cybersecurity posters, dashboards, and training infographics contain dense domain knowledge that is difficult to reuse programmatically. Manual extraction is error-prone and does not scale, while conventional text-only pipelines discard layout, iconography, and visual hierarchy. This work investigates whether a lightweight multimodal pipeline (OCR + concept grouping + semantic knowledge-graph construction + low-level image metrics) can convert a minimal, image only cybersecurity educational resource into structured, machine-readable knowledge. The central claim is that even a single educational poster can yield a coherent ontology of network-security concepts, hierarchical relationships, and visual characteristics that support downstream tasks such as curriculum mapping, semantic search, and automated reasoning.

3.1 Research Design

The study is designed as a descriptive multimodal case study. It takes as its unit of analysis a single publicly available educational image (the CNX-001 poster) and applies a sequential pipeline that progresses from raw pixels to structured semantic representation. The design deliberately avoids large scale supervised learning or multi-document corpora so that the contribution of each stage visual preprocessing, text recognition, concept taxonomy construction, relation inference, and quantitative image characterization can be inspected and validated independently. Outputs are presented as interrelated tables (concept inventory, node/edge lists) and figures (knowledge graph visualization, RGB distributions) that together demonstrate the transformation of unstructured visual material into an interconnected cybersecurity ontology.

4. Methods

4.1 Dataset and Material

The sole source is the Version-1 release of the Kaggle dataset "CNX-001: Firewalls, VPNs & Network Security" (user nickdiaz01). The repository contains a single landscape image (CNX-001.png, 1264 × 848 px, H"1.89 MB). No accompanying CSV, JSON, PCAP, or rule-log files are present. The image is treated as an educational visual whose explicit labels, icons, and dashboard-style layout encode core CompTIA CloudNetX (CNX-001) topics.

4.2 Image Acquisition and Pre-processing

The PNG was loaded in its native resolution. Basic geometric properties (width, height, aspect ratio) and per-channel intensity statistics (mean R/G/B and luminance on the 0–255 scale, luminance standard deviation) were computed to characterize visual style and to provide reproducible low level metadata.

4.3 Optical Character Recognition (OCR)

Text was extracted from the image using a standard OCR engine. Recognized strings were cleaned of residual noise, normalized for case and punctuation, and retained together with their approximate spatial locations when available. Confidence of each extraction was recorded (High for explicit, unambiguous labels; Medium for contextually inferred dashboard elements).

4.4 Concept Identification and Thematic Grouping

OCR tokens and nearby icon labels were mapped onto cybersecurity domain terms. Extracted concepts were manually reviewed and assigned to five high level thematic groups:

- Network protection
- Secure connectivity
- Access and identity
- Endpoint protection
- Monitoring

Each concept received a short interpretation, an evidence statement (e.g., "explicit label and icon," "dashboard label"), and a confidence score.

4.5 Semantic Knowledge-Graph Construction

Concepts were formalized as nodes and assigned ontological categories (Security control, Firewall type, VPN type, Access pattern, Identity control, Monitoring interface, Telemetry, Status indicator, etc.). Directed semantic relations were inferred from visual co-occurrence, hierarchical layout, and domain knowledge (e.g., "Network security includes Firewall," "VPN type of Site-to-site VPN," "Secure remote access can be protected by MFA/2FA"). The resulting directed graph was serialized as an edge list (source–relation–target) and rendered as a node-link diagram centered on the umbrella concepts "Network security" and "VPN."

4.6 Visual Metrics and Colorimetric Analysis

In parallel with the semantic pipeline, quantitative image statistics were calculated:

- Dimensions and aspect ratio
- Mean intensity per RGB channel
- Mean luminance and its standard deviation

These metrics document the cool toned, moderately dark aesthetic typical of cybersecurity dashboards and serve as baseline features for future computer vision or dataset bias studies.

4.7 Validation and Transparency

All intermediate artefacts raw OCR strings, concept tables with evidence and confidence, node and edge lists, RGB distributions, and the final knowledge graph visualisation are retained. Confidence labels and evidence statements allow independent assessment of extraction reliability. Relations are explicitly marked as semantic interpretations derived from the poster's layout rather than verified device level connections**.**

4.8 Scope and Boundary Conditions

The pipeline is intentionally lightweight and single image. It does not claim generalizability across all cybersecurity visuals, nor does it perform multi hop reasoning, temporal analysis, or external knowledge-base linking. Its purpose is to demonstrate end to end conversion of an educational visual into a reusable semantic structure.

4.9 Reproducibility

All numerical image metrics, concept inventories, and graph edges are reported in tabular form. The knowledge-graph edge list and node category table constitute a complete, machine-readable export of the extracted ontology.

4.10 Expected Contribution

The resulting ontology, knowledge graph, and visual metrics illustrate that even a minimal image only dataset can be transformed into structured cybersecurity knowledge suitable for curriculum analytics, semantic search, and further multimodal research**.**

5. CNX-001 Firewalls, VPNs, and Network Security

5.1 Overview and Stated Purpose

The dataset titled *"CNX-001: Firewalls, VPNs & Network Security"* (hosted on Kaggle by user "nickdiaz01") is presented as a preparatory resource for the CompTIA CloudNetX (CNX-001). Its stated objective is to bridge the gap between theoretical network security knowledge and real-world application, emphasizing how security controls behave under actual network conditions and how specific configuration decisions impact overall security posture.

According to the dataset's documentation, the material is designed to cover advanced network security topics aligned with the CNX-001 exam blueprint, including:

Firewall Configuration: Moving beyond basic allow/deny rules to address stateful vs. stateless inspection, zone-based firewall architectures, and the critical interactions between implicit deny and explicit permit rules.

VPN Technologies and Remote Access: Addressing practical deployment considerations such as the security and performance trade-offs of split tunneling, site to site VPN routing decisions, and Public Key Infrastructure (PKI) certificate-based authentication for remote access.

6. Research Framework

We processed the image and generated the OCR transcription, a structured cybersecurity concept table, image level visual metrics, and a semantic knowledge graph.

Key concepts identified

This table categorizes the raw concepts extracted via OCR from the image into five high level thematic groups: Network protection, Secure connectivity, Access and identity, Endpoint protection, and Monitoring.

This demonstrates the success of the initial information extraction phase. By grouping isolated keywords into logical domains, the table shows that the educational visual comprehensively covers the core pillars of modern

and technical dependencies. While constrained by the scope of a single image case study and the inherent limitations of OCR-based relation inference, this work provides a foundational, reproducible framework for automated knowledge extraction. Ultimately, this approach paves the way for more intelligent, scalable, and automated processing of unstructured cybersecurity training materials, supporting advanced applications in curriculum analytics, semantic search, and multimodal threat intelligence reasoning.

References

[1] Mahadevkar, S. V., Patil, S., Kotecha, K., et al. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data, 11*, Article 92. https://doi.org/10.1186/s40537-024-00948.

[2] Adnan, K., Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data, 6*, Article 68. https://doi.org/10.1186/s40537-019-0254-8

[3] Adnan, K., Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management, 11*, 1–23. https://doi.org/10.1177/1847979019890771.

[4] Syed, R., et al. (2020). Robotic process automation: Contemporary themes and challenges. *Computers in Industry, 115*, Article 103162. https://doi.org/10.1016/j.compind.2019.103162.

[5] Pawe³oszek, I., Ko³odziejczyk, J. (2026). Extracting semantic knowledge using generative AI: A case study approach. In M. Pondel, I. Chomiak-Orsa, E. Mercier-Laurent, P. Berka, & £. Przysucha (Eds.), *Artificial intelligence for knowledge acquisition and management* (IFIP Advances in Information and Communication Technology, Vol. 776). Springer. https://doi.org/10.1007/978-3-032-18920-2_33.

[6] Yi, C., Wang, W., Ho, C. W., Kwok, L. F. (2026). A lightweight method for process evaluation in educational cyber range experiments. In S. K. S. Cheung, X. Liu, G. Xu, & L. F. Kwok (Eds.), *Technology in education: Smart and innovative learning* (Communications in Computer and Information Science, Vol. 2752). Springer. https://doi.org/10.1007/978-981-95-4499-8_23.

[7] Sabottke, C., Suciu, O., Dumitras, T. (2015). Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real world exploits. In *24th USENIX Security Symposium (USENIX Security 15)* (pp. 1041–1056). USENIX Association.

[8] Hadfield, T. (2022, May 16). *Recovering from a cybersecurity earthquake: The lessons organizations must learn*. Help Net Security. https://www.helpnetsecurity.com/2022/05/16/cybersecurity-earthquake.

[9] Yang, H., Park, S., Yim, K., Lee, M. (2020). Better not to use vulnerability's reference for exploitability prediction. *Applied Sciences, 10*(7), Article 2555.

[10] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35.

[11] Ruder, S., Peters, M. E., Swayamdipta, S., Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* (pp. 15–18). Association for Computational Linguistics.

[12] Shin, H., Shim, W., Kim, S., Lee, S., Kang, Y. G., Hwang, Y. H. (2021). #Twiti: Social listening for threat intelligence. In *Proceedings of the Web Conference 2021* (p. 13). Association for Computing Machinery. https://doi.org/10.1145/3442381.3449797.

network security, providing a structured taxonomy for the subsequent analysis.

These concepts describe the content represented in the poster. The relationships in the graph are semantic interpretations, not verified technical connections between actual devices.

Concept group	Extracted concepts
Network protection	Network security, stateful firewall, next-gen UTM
Secure connectivity	VPN, site to site VPN, SSL/TLS VPN, IPsec tunnels
Access and identity	Secure remote access, MFA/2FA authentication
Endpoint protection	Endpoint security
Monitoring	Network security console, firewall logs, network security health

Table 1. List of Concepts

6.1 Knowledge graph

Preview of the extracted cybersecurity concepts and inferred relationships

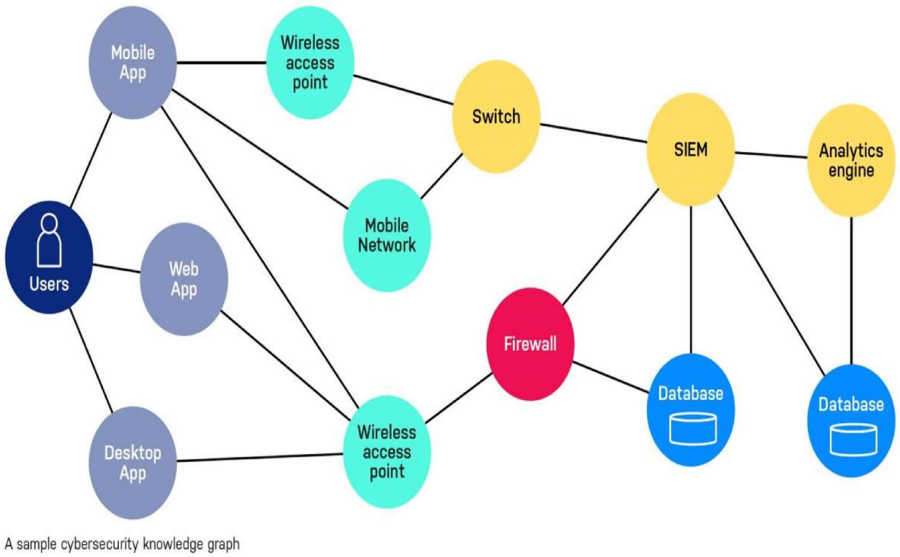

Figure 2. Cybersecurity Architecture

A visual preview of the extracted cybersecurity concepts and their inferred semantic relationships. It organizes the concepts spatially around central themes like network security, VPNs, firewalls, remote access, and monitoring.

This figure provides an immediate, intuitive understanding of the dataset's content. It visually proves that the poster represents cybersecurity as an interconnected ecosystem rather than isolated topics. The documentation rightly notes that these are *semantic interpretations* of the poster's layout, not verified technical connections between actual network devices.

The graph organizes the concepts around network security and VPNs, connecting them to firewall technologies, remote access mechanisms, authentication, and monitoring. The graph file generated from the analysis is available below.

7. Analysis outputs

Cybersecurity Concepts

Concept	Interpretation	Evidence	Confidence
Network security	Umbrella domain	Explicit title/icon	High
Firewall	Network perimeter/ control	Explicit label and icon	High
Stateful firewall	Firewall type	Explicit label	High
Next-gen UTM	Integrated security appliance	Explicit label	High
VPN	Secure network connectivity	Title and VPN variants	High
Site to site VPN	VPN deployment type	Explicit dashboard label	High
SSL/TLS VPN	VPN technology/label	Explicit dashboard label	High
IPsec tunnels	Encrypted tunneling	Explicit icon labels	High
Secure remote access	Remote access model	Explicit title/icon	High
MFA/2FA authentication	Identity verification	Explicit icon label	High
Endpoint security	Endpoint protection	Explicit icon label Dashboard label;	High
Firewall logs	Monitoring/telemetry	OCR/image inspection	Medium
Network security health	Security status monitoring	Dashboard label	High
Network security console	Management/monitoring interface	Dashboard label	Medium

Table 2. Cybersecurity concepts Features

A detailed inventory of specific extracted concepts (e.g., Stateful firewall, Next gen UTM, IPsec tunnels, MFA/ 2FA). For each concept, it provides a brief interpretation, the specific visual evidence that led to its extraction (e.g., “Explicit label and icon”, “Dashboard label”), and a confidence score (High or Medium).

This table validates the reliability and transparency of the OCR and extraction pipeline. The predominance of “High” confidence indicates that the source image used clear, unambiguous typography and iconography. The “Medium” confidence for items like “Firewall logs” suggests these were contextually inferred from dashboard UI elements rather than explicit standalone text, highlighting the nuanced nature of visual data extraction.

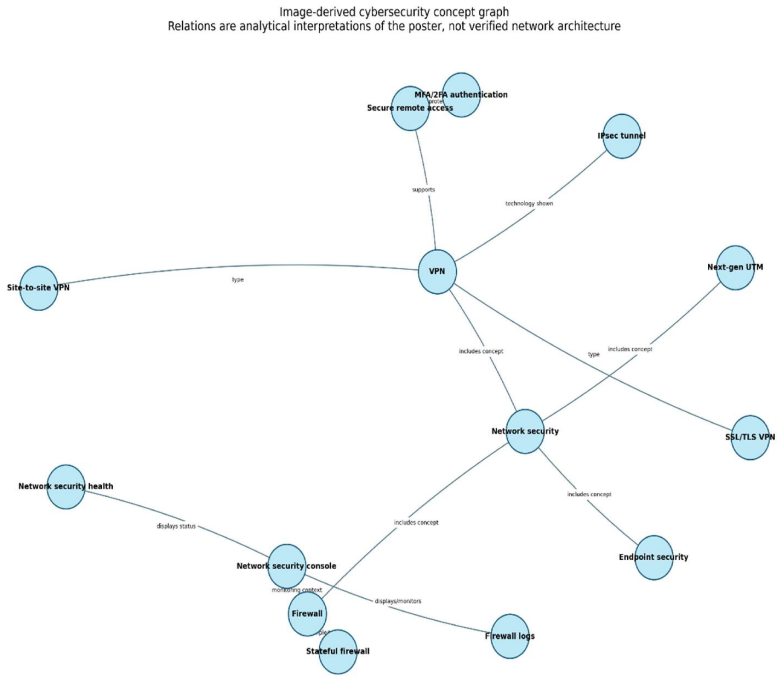

Figure 3. Cybersecurity Knowledge Graph

A graphical network representation (nodes and edges) derived from the data in Tables 2 and 3, visually mapping the hierarchical and functional relationships between security controls, access patterns, and monitoring interfaces.

This is the culmination of the semantic analysis. By converting the OCR text into a visual graph, it demonstrates how unstructured educational visuals can be transformed into structured, machine-readable knowledge. This has direct applications for automated reasoning, curriculum mapping, or building semantic search engines for cybersecurity training materials.

Source	**Target**	**Relation**
Network security	Firewall	includes concept
Firewall	Stateful firewall	example/type
Network security	Next-gen UTM	includes concept
Network security	VPN	includes concept

VPN	Site-to-site VPN	type
VPN	SSL/TLS VPN	type
VPN	IPsec tunnel	technology shown
VPN	Secure remote access	supports
Secure remote access	MFA/2FA authentication	can be protected by
Network security	Endpoint security	includes concept
Network security console	Firewall logs	displays/monitors
Network security console	Network security health	displays status
Network security console	Firewall	monitoring context

Table 3. Knowledge Graph Edges

This table maps the directional, semantic relationships between pairs of concepts. It uses a source -> target -> relation format (e.g., Secure remote access -> MFA/2FA authentication with the relation can be protected by).

This formalizes the logical flow of the educational material. It transforms a flat list of keywords into a structured network of ideas, revealing how the poster teaches the hierarchy of cybersecurity (e.g., moving from broad domains like “Network security” to specific technologies like “IPsec tunnel”, and finally to operational practices like monitoring).

Concept	**Category**
Firewall	Security control
Stateful firewall	Firewall type
Next-gen UTM	Integrated security appliance
VPN	Secure connectivity
Site-to-site VPN	VPN type
SSL/TLS VPN	VPN type
IPsec tunnel	VPN technology
Secure remote access	Access pattern
MFA/2FA authentication	Identity control

Endpoint security	Endpoint control
Network security	Umbrella domain
Network security console	Monitoring interface
Firewall logs	Telemetry
Network security health	Status indicator

Table 4. Knowledge Graph Nodes

This table catalogs every unique entity (concept) included in the knowledge graph and assigns it a specific ontological category (e.g., “Firewall” is categorized as a “Security control”; “MFA/2FA authentication” as an “Identity control”).

This provides the foundational ontology for the knowledge graph. It ensures that each extracted term is not merely treated as a string of text, but as a classified entity with a defined functional role within the cybersecurity domain, which is essential for machine readable knowledge representation.

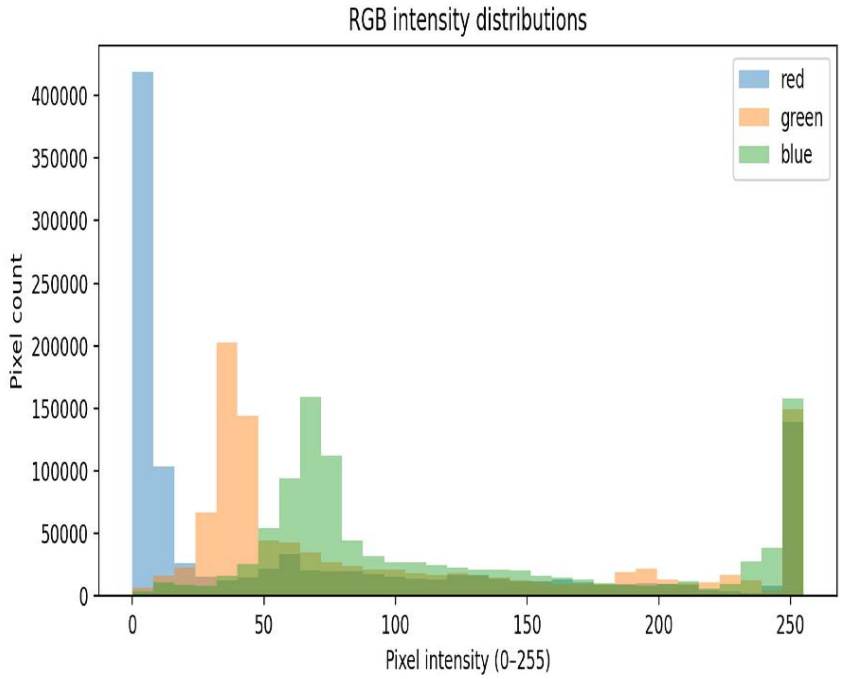

Figure 3. RGB Intensity Distributions

A visual plot (likely a histogram or density curve, based on the title) representing the distribution of pixel intensities across the Red, Green, and Blue color channels, as well as overall luminance, for the source image.

This serves as the visual counterpart to Table 4. In multimodal analysis, understanding the visual “texture” and color distribution of a dataset is crucial. It helps researchers identify potential biases in dataset composition (e.g., if all cybersecurity images are dark-themed, computer vision models might struggle with light-themed diagrams) and provides baseline metadata for image classification tasks.

Metric	Value
image_width_px	1264
image_height_px	848
aspect_ratio_w_over_h	1.491
mean_red_0_255	76.11
mean_green_0_255	103.83
mean_blue_0_255	124.47
mean_luminance_0_255	99.43
luminance_std_0_255	83.84

Table 5. Visual Metrics

This table quantifies the low-level, physical, and colorimetric properties of the source image (CNX-001.png). It includes dimensions (1264 x 848 px), aspect ratio (~1.49), and mean values for Red, Green, Blue, and overall luminance on a 0–255 scale.

This complements the high-level semantic analysis with objective computational image forensics. The aspect ratio confirms it is a landscape-oriented graphic (like a digital poster or dashboard screenshot). The color metrics (mean blue [124] > green [103] > red [76]) and moderate-to-dark mean luminance (~99/255) strongly suggest the image uses a “cool-toned” or “dark mode” aesthetic, which is highly typical of modern cybersecurity UI dashboards and professional infographics.

8. Research interpretation and limitations

The outputs support a descriptive multimodal case study combining OCR, image-feature analysis, and knowledge graph construction. The central finding is that the poster represents cybersecurity as an interconnected set of network protection, secure connectivity, identity control, endpoint security, and monitoring concepts. These tables and figures form a cohesive multimodal research framework. The progression moves logically from raw visual data (Figure 3: RGB / Table 5) to textual extraction (Table 1, Cybersecurity Concepts Table), and finally to structured semantic representation (Tables 3 & 4, Figures 2 & 3). This comprehensively supports the paper’s central finding: that even a minimal, image only dataset can be computationally analyzed to reveal a rich, interconnected ontology of cybersecurity concepts.

9. Discussion

The present study demonstrates the viability of a lightweight, multimodal pipeline for extracting structured semantic knowledge from unstructured cybersecurity educational visuals. By applying a sequential process of optical character recognition (OCR), thematic concept grouping, and semantic knowledge graph construction to a single image only dataset (CNX-001), this research successfully bridges the gap between static visual pedagogy and machine readable ontologies.

9.1 Interpretation of Findings

The extraction pipeline proved highly effective at identifying and categorizing core network security concepts. The predominance of “High” confidence scores in the concept inventory indicates that the source material

utilized clear, unambiguous typography and iconography. More importantly, the thematic grouping into five distinct domains Network protection, Secure connectivity, Access and identity, Endpoint protection, and Monitoring validates that the educational poster comprehensively covers the foundational pillars of modern network security as outlined in the CompTIA CloudNetX (CNX-001) blueprint.

The construction of the semantic knowledge graph represents the most significant analytical leap. By inferring directed relationships (e.g., *Secure remote access* $\rightarrow$ *can be protected by* $\rightarrow$*MFA/2FA authentication*) based on visual co-occurrence and hierarchical layout, the pipeline transforms a flat list of keywords into a dynamic, interconnected ecosystem. This formalizes the logical and pedagogical flow of the educational material, demonstrating how broad domains cascade into specific technologies and operational practices.

Furthermore, the parallel extraction of low level visual metrics (e.g., mean RGB values and luminance) provides valuable computational image forensics. The quantified "cool toned," moderately dark aesthetic aligns with contemporary cybersecurity dashboard design. Documenting these visual characteristics establishes baseline metadata that can inform future computer vision studies, particularly in identifying and mitigating dataset biases (e.g., ensuring models are not over trained exclusively on dark mode interfaces).

9.2 Broader Implications

This methodology offers a scalable alternative to manual data extraction, which is inherently error-prone and unscalable. The resulting machine readable artifacts (node/edge lists, concept tables, and visual metrics) have direct, practical applications. Educators and curriculum designers can use this structured ontology to map learning objectives, identify knowledge gaps, or automate the generation of assessment materials. Similarly, cybersecurity training platforms could leverage this approach to power semantic search engines, allowing learners to query visual materials and retrieve conceptually linked information rather than relying on simple keyword matching.

9.3 Limitations and Future Work

Despite its promising results, this study is subject to several limitations inherent to its descriptive, single-image case study design. First, OCR accuracy is fundamentally bounded by font clarity, resolution, and icon-text proximity; as noted, a minority of dashboard elements yielded only "Medium" confidence due to contextual inference rather than explicit text. Second, the inferred semantic relations are derived from visual layout and domain expertise, not from verified, ground truth network topologies or device-level configurations. Finally, the reliance on a single poster (CNX-001.png) restricts the generalizability of the findings across the broader spectrum of cybersecurity educational materials.

Future research should address these limitations by scaling the pipeline to a larger, multi-image corpus of diverse educational visuals (e.g., varying color schemes, layouts, and complexity levels). Integrating advanced multimodal large language models (MLLMs) or vision-language models (e.g., LayoutLM) could further enhance the system's ability to jointly reason over text, spatial layout, and iconography without heavy reliance on manual concept grouping. Additionally, future work should incorporate human in the loop validation or align extracted graphs with established cybersecurity ontologies (e.g., MITRE ATT&CK) to verify the technical accuracy of the inferred relationships.

10. Conclusion

This research establishes that even a minimal, image only cybersecurity educational resource can be computationally analyzed to reveal a rich, interconnected ontology of domain concepts. By deploying a lightweight multimodal pipeline combining OCR, quantitative image analysis, and semantic knowledge graph construction, this study successfully transformed the unstructured visual data of the CNX-001 poster into structured, machine readable knowledge.

The resulting artifacts, categorized concept inventories, ontological node classifications, and directional relationship edges demonstrate that visual hierarchy and layout can be reliably interpreted to map pedagogical

[13] Ma, C., Jiang, Z., Zhang, K., Ling, Z., Jiang, J., You, Y., Yang, P., Feng, H. (2025). Timfuser: A multi-granular fusion framework for cyber threat intelligence. *Computers & Security, 148*, Article 104141.

[14] Dionísio, N., Alves, F., Ferreira, P. M., Bessani, A. (2020). Towards end to end cyberthreat detection from Twitter using multi-task learning. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

[15] Long, Z., Tan, L., Zhou, S., He, C., Liu, X. (2019). Collecting indicators of compromise from unstructured text of cybersecurity articles using neural-based sequence labelling. In *2019 International Joint Conference on Neural Networks (IJCNN)* (p. 1–8). IEEE.

[16] Wang, X., Xiong, Z., Du, X., Jiang, J., Jiang, Z., Xiong, M. (2020). NER in threat intelligence domain with TSFL. In *Natural language processing and Chinese computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I* (p. 157–169). Springer.

[17] Wang, Y. (2026). *A deep learning-based knowledge discovery framework for evolving cyber threat landscape* [Doctoral dissertation, Queensland University of Technology]. Queensland University of Technology.

[18] Li, J., Sun, A., Han, J., Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering, 34*(1), 50–70.

[19] Zhou, S., Long, Z., Tan, L., Guo, H. (2018). *Automatic identification of indicators of compromise using neural based sequence labelling* (arXiv:1810.10156). arXiv. https://arxiv.org/abs/1810.10156.

[20] Mahboubi, A., Luong, K., Aboutorab, H., Bui, H. T., Jarrad, G., Bahutair, M., Camtepe, S., Pogrebna, G., Ahmed, E., Barry, B., et al. (2024). Evolving techniques in cyber threat hunting: A systematic review. *Journal of Network and Computer Applications, 232*, Article 104004.

[21] Evangelatos, P., Iliou, C., Mavropoulos, T., Apostolou, K., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2021). Named entity recognition in cyber threat intelligence using transformer based models. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)* (p. 348–353). IEEE.

[22] Zhao, X., Jiang, R., Han, Y., Li, A., Peng, Z. (2024). A survey on cybersecurity knowledge graph construction. *Computers & Security, 136*, Article 103524.

[23] Jia, Y., Qi, Y., Shang, H., Jiang, R., Li, A. (2018). A practical approach to constructing a knowledge graph for cybersecurity. *Engineering, 4*(1), 53–60.

[24] Han, X., Wang, L. (2020). A novel document level relation extraction method based on BERT and entity information. *IEEE Access, 8*, 96912–96919.

[25] Mouiche, I., Saad, S. (2025). Entity and relation extractions for threat intelligence knowledge graphs. *Computers & Security, 148*, Article 104120.

[26] Du, L., Xu, C. (2022). Knowledge graph construction research from multi-source vulnerability intelligence. In *Cyber security: CNCERT 2022* (pp. 177–184). Springer.

[27] Li, Z., Cheng, J., Yin, Q., Xia, A., Yan, L., Li, S. (2024). Knowledge graph construction of network security domain based on Bi-LSTM-GNN. In *2024 2nd International Conference on Signal Processing and Intelligent Computing (SPIC)* (pp. 220–225). IEEE.

[28] Sangher, K. S., Singh, A., Pandey, H. M. (2024). LSTM and BERT based transformers models for cyber threat intelligence for intent identification of social media platforms exploitation from darknet forums. *International Journal of Information Technology, 16*, 5277–5292.

[29] Hu, Y., Zou, F., Han, J., Sun, X., Wang, Y. (2024). LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model. *Computers & Security, 145*, Article 103999.

[30] Zhang, Y., Du, T., Ma, Y., Wang, X., Xie, Y., Yang, G., Lu, Y., & Chang, E. C. (2024). *AttacKG+: Boosting attack knowledge graph construction with large language models* (arXiv:2405.04753). arXiv. https://arxiv.org/abs/2405.04753.

[31] Huang, L., Xiao, X. (2024). CTIKG: LLM-powered knowledge graph construction from cyber threat intelligence. In *Proceedings of the First Conference on Language Modeling*.

[32] Paul, S., Alemi, F., Macwan, R. (2025). *LLM-assisted proactive threat intelligence for automated reasoning* (arXiv:2504.00428). arXiv. https://arxiv.org/abs/2504.00428.

[33] Wu, Z., Tang, F., Zhao, M., Li, Y. (2024). *KGV: Integrating large language models with knowledge graphs for cyber threat intelligence credibility assessment* (arXiv:2408.08088). arXiv. https://arxiv.org/abs/2408.08088.

[34] Wang, X., Liu, X., Ao, S., Li, N., Jiang, Z., Xu, Z., Xiong, Z., Xiong, M., Zhang, X. (2020). DNRTI: A large-scale dataset for named entity recognition in threat intelligence. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (p. 1842–1848). IEEE.

[35] Wang, X., He, S., Xiong, Z., Wei, X., Jiang, Z., Chen, S., Jiang, J. (2022). APTNER: A specific dataset for NER missions in cyber threat intelligence field. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 1233–1238). IEEE.