LLM-Based Classification of CVEs for Vulnerability Analysis in Medical IT Systems

Pere Vidiella , Pere Tuset-Peiró * , Josep Pegueroles , Michael Pilgermann

Posted Date: 10 March 2026

doi: 10.20944/preprints202603.0737.v1

Keywords: medical IT systems; LLM; CVE; CVSS; PACS; DICOM; HL-7; MITRE ATT&CK

 doi:10.20944/preprints202603.0737.v1

Article

LLM-Based Classification of CVEs for Vulnerability Analysis in Medical IT Systems

Pere Vidiella [1,2], Pere Tuset-Peiró [2,3,*], Josep Pegueroles [1] and Michael Pilgermann [4]

[1] Universitat Politècnica de Catalunya, Barcelona, Spain
[2] TecnoCampus - Universitat Pompeu Fabra, Mataró, Spain
[3] Universitat Oberta de Catalunya, Barcelona, Spain
[4] Brandenburg University of Applied Sciences, Brandenburg, Germany
* Correspondence: ptuset@tecnocampus.cat

Abstract

The digitalization of healthcare systems increases their exposure to security incidents. Security analysts use standard CVE (Common Vulnerabilities and Exposures) records to identify and mitigate vulnerabilities. However, CVEs are often incomplete or overly generic, requiring the addition of structured, actionable information to support effective decision-making. Manually performing this augmentation is unfeasible due to the rapidly growing number of published CVEs. In this paper we evaluate the capabilities of LLMs (Large Language Models) to classify and analyze CVEs within the medical IT systems domain. We propose a framework where LLMs parse structured JSON context and answer a set of specific natural language questions, enabling the categorization of vulnerabilities by their position in the medical chain, affected component types, and mapping to the MITRE ATT&CK framework. While recent studies show that general LLMs can achieve high accuracy in objective CVSS elements and learn CNA-oriented patterns, they often struggle with subjective impact metrics. Our results demonstrate that domain-specific classification through natural language prompting provides the necessary granularity for medical risk prioritization. We conclude that this augmentation effectively bridges the gap in standard CVE records, allowing for a better understanding of how vulnerabilities impact critical healthcare infrastructure and patient safety.

Keywords: medical IT systems; LLM; CVE; CVSS; PACS; DICOM; HL-7; MITRE ATT&CK

Tipo de contribución: *Investigación en desarrollo*

1. Introduction

The digital transformation of healthcare has led to an unprecedented reliance on medical IT systems, which are now the backbone of patient care, diagnostic workflows, and hospital management. Ensuring the confidentiality, integrity, and availability of these systems is not merely a technical requirement but a fundamental necessity for service continuity and patient safety. A breach in confidentiality can expose sensitive personal health information, while a compromise in integrity or availability can lead to incorrect diagnoses or the catastrophic disruption of life-critical medical services. Central to managing these risks is the CVE (Common Vulnerabilities and Exposures) system [1], which provides a standardized framework for identifying and cataloging publicly known security vulnerabilities. However, as the volume of reported CVEs grows rapidly—increasing by 38% between 2023 and 2024 alone—the burden on cybersecurity researchers and practitioners to manually analyze and prioritize these threats has become unsustainable, particularly in the complex, heterogeneous environments of medical IT.

Recent research has begun to explore the role of LLMs (Large Language Models) for various network-related tasks [2], including automating the assessment of vulnerabilities. For example, Marchiori et al. [3] investigates the capabilities of various LLMs, such as Gemma 3 and GPT-4o Mini, in generating CVSS (Common Vulnerability Scoring System) vectors directly from CVE descriptions.

 doi:10.20944/preprints202603.0737.v1

Their findings suggest that while LLMs excel at identifying objective metrics like the AV (Attack Vector), they struggle with more subjective components like impact metrics, where traditional embedding-based models often perform better. Complementary to this, Miranda et al. [4] focuses on the influence of CNAs (CVE Numbering Authorities), demonstrating that LLMs like CVSS-BERT can learn latent patterns to determine how different vendors and organizations impact vulnerability severity scores based on their specific product context. More recently, Mirtaheri et al. [5] introduce a knowledge-driven approach that anchors LLM reasoning in structured KGs (Knowledge Graphs). By mapping CVSS definitions into decision trees and injecting this expert logic into the prompt, they achieved accuracy improvements from 59% to over 82% using GPT-4o, effectively bridging the gap between symbolic reasoning and statistical generation. To further address the scalability of these systems, Jafarikhah et al. [6] evaluates advanced models, including GPT-5 and Gemini 2.5, across 31,000 CVEs. Their work highlights that while ensemble-based "meta-classifiers" can marginally improve scoring precision, the primary bottleneck remains the ambiguity of vulnerability descriptions rather than model capacity. Overall, these studies underscore the potential of LLMs to streamline the scoring process but emphasize the need for context-aware, structured, and agentic workflows to achieve reliable results in sensitive domains.

While in previous works we focused on assessing the state of cybersecurity in medical IT systems using OSINT tools [7], in this paper we introduce a novel approach for CVE vulnerability classification and analysis specifically tailored to the medical IT domain. While previous works have focused on general CVSS scoring or CNA-oriented patterns, our research leverages LLMs to perform a deep, multi-dimensional augmentation of CVE data using natural language classification that enables enhanced vulnerability classification and analysis. By presenting the LLM with CVE context in a structured JSON format, we enable the model to:

1. Filter for domain-specific relevance within medical IT systems.
2. Precisely identify affected vendors and products to quantify vulnerability trends.
3. Categorize vulnerabilities according to their position in the medical chain, from data generation to interoperability systems.
4. Determine the specific component type and map the CVE to relevant MITRE ATT&CK categories.

This specialized focus provides a level of granularity and interpretability that is typically absent from standard CVE records, offering a first-of-its-kind framework for automatically augmenting vulnerability intelligence to better protect the critical infrastructure of modern healthcare.

The remainder of the paper is organized as follows. Section 2 presents the essential cybersecurity concepts that are used throughout the paper. Section 3 presents an overview of medical IT systems that is later used for vulnerability analysis. Section 4 describes the process to obtain the results, from creating the CVE database to processing it using LLMs. Section 5 outlines and discusses the results using the presented methodology. Finally, Section 6 summarizes our findings and outlines the future work.

2. Overview of Cybersecurity Concepts

In this section we present an overview of the essential concepts that are used for vulnerability and adversary tactics classification in the field of cybersecurity: CVE catalog and MITRE ATT&CK framework.

2.1. CVE Catalog

CVE is an industry-standard catalog of publicly known information cybersecurity vulnerabilities and exposures. It was funded by the U.S. Department of Homeland Security and is currently operated by the MITRE Corporation. It works by assigning unique alphanumeric identifier to specific vulnerabilities, allowing security automation tools, patch management systems, and researchers to accurately correlate and share information regarding specific threats across disparate platforms. Among others, CVE entries contain the following information:

 doi:10.20944/preprints202603.0737.v1

- **CVE identifier**: A unique, standardized reference key (e.g., `CVE-YYYY-NNNNN`) assigned to a specific vulnerability for cross-database correlation.
- **Description**: A text summary explaining the vulnerability's root cause, attack vector, and potential impact.
- **Affected system**: The specific vendor, product, and version(s) in which the vulnerability exists.
- **Severity metrics (CVSS)**: A standardized framework that assigns a quantitative base score (0.0 to 10.0) reflecting the flaw's intrinsic severity. This score is derived from exploitability metrics—such as the required attack vector, complexity, and privilege levels—as well as the potential impact.
- **Administrative metadata**: Details regarding the assigning CNA (CVE Numbering Authority) and timestamps for the record's publication and most recent updates.

2.2. MITRE ATT&CK Framework

The MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) framework is a globally accessible, widely adopted knowledge base and model for cybersecurity adversary behavior. Developed by the MITRE corporation, it is built upon empirical observations of real-world cyber-attacks, and provides a standardized taxonomy of the operational phases an adversary undertakes during an intrusion. The MITRE ATT&CK facilitates the systematic categorization of threat actor behaviors, enabling organizations to map specific vulnerabilities to known adversarial methodologies. The framework organizes adversary behavior into a matrix structured around three primary hierarchical categories:

- **Tactics**: Represent the specific, short-term operational objectives an adversary seeks to achieve during the life-cycle of a cyberattack. The Enterprise matrix outlines 14 distinct tactics that encompass the entire attack continuum. These span from initial network compromise (Initial Access, Execution), to maintaining presence and stealth (Persistence, Privilege Escalation, Defense Evasion, Credential Access), navigating the environment (Discovery, Lateral Movement), and finally executing ultimate objectives against the target data or systems (Collection, Command and Control, Exfiltration, Impact).
- **Techniques**: Detail the precise technological or operational methods employed by an adversary to accomplish a given tactical objective. For example, to achieve Initial Access, an attacker might utilize the technique of Phishing or exploiting public-facing applications.
- **Procedures**: Document the exact operational steps, software tools, and malware signatures utilized by specific threat groups (e.g., Advanced Persistent Threats, or APTs) to execute a technique. For example, a given procedure might detail how a specific ransomware syndicate exploits a known vulnerability to encrypt file-systems.

3. Overview of Medical IT Systems

In this section we provide an overview of the IT systems used within the medical domain, classified according to the tasks they perform. Such classification has been derived from established healthcare information system architectures and standards, including HL7 [8], ISO/HL7 10781 [9], ISO 12967 [10], and DICOM [11], as well as from widely cited biomedical informatics literature [12,13]. In particular, we identify the following system categories:

1. **Data identity and management systems**: These systems form the administrative backbone of healthcare organizations, managing patient identity and life-cycle. These systems include:
    - **HIS (Hospital Information System)**: Designed to manage medical, administrative, financial, and legal aspects of hospital operations, serving as the central integration point for departmental subsystems.
    - **EMR (Electronic Medical Record)**: Digital versions of traditional paper-based medical charts. EMRs are typically confined to a single healthcare provider and are not designed for extensive data sharing.

 doi:10.20944/preprints202603.0737.v1

- **EHR (Electronic Health Record)**: Electronic repositories of patient health information designed to support data sharing across multiple healthcare providers and organizations.

2. **Departmental information systems**: These systems support the operational workflows of specific clinical departments inside a medical organization, managing specialized ordering, tracking, and result-reporting processes, and acting as an intermediate operational layer between clinicians and the core medical records. These systems include:
   - **RIS (Radiology Information System)**: Systems for managing radiological workflows, often integrated with PACS and VNA solutions.
   - **LIS (Laboratory Information System)**: Managing laboratory operations such as order entry, specimen tracking, result reporting, and quality control.
   - **PIS (Pharmacy Information System)**: Supporting medication management, including drug dispensing, inventory control, and clinical checks.
3. **Data generation systems**: Data generation systems comprise medical equipment and devices that produce clinical data, including diagnostic imaging modalities and bedside devices. These devices are typically integrated with medical IT systems via interoperability standards, such as DICOM or HL7. Examples of data generation systems include: CT (Computed Tomography), MRI (Magnetic Resonance Imaging), PET (Positron Emission Tomography), and X-Ray, among others.
4. **Storage and archiving systems**: These systems provide long-term storage and access to high-volume clinical data, particularly medical images, ensuring data integrity and availability. These systems include:
   - **PACS (Picture Archiving and Communication System)**: Providing efficient storage, retrieval, and distribution of medical images generated by multiple imaging modalities.
   - **VNA (Vendor Neutral Archive)**: Centralized archives that store medical images and documents in standardized formats.
5. **Interoperability standards**: Define data communication and storage protocols that act as middleware to enable data exchange between heterogeneous healthcare systems. These standards include:
   - **DICOM (Digital Imaging and Communications in Medicine)**: International standard for the secure transmission, storage, and sharing of medical images (like MRIs or CT scans) and their associated data. It ensures interoperability between imaging equipment and healthcare IT systems across different manufacturers.
   - **HL7 (Health Level Seven)**: A framework of international standards governing the exchange, integration, and retrieval of electronic health information. It acts as a universal language that allows disparate medical software systems to share clinical and administrative data seamlessly.
6. **IT communication systems**: The underlying network infrastructure that interconnects medical IT and clinical systems, enabling data exchange and service availability. They include wired and wireless networks, segmentation mechanisms such as VLANs, and security controls such as firewalls and access policies, among others.

The proposed classification provides a functional context for CVEs by mapping individual vulnerabilities to the medical IT systems in which they are likely to manifest. While CVE records describe technical flaws in isolation, this system-level classification enables the interpretation of vulnerabilities in terms of their operational and clinical impact within healthcare environments. In addition, while IT communication systems are extensively used in the medical context to provide connectivity between medical IT systems, we do not consider them in the remainder of the paper because the vulnerabilities are not exclusive of medical IT systems.

 doi:10.20944/preprints202603.0737.v1

4. Methodology

In this section we present the different steps that we have performed to create the CVE database, prepare the LLM setup, define the tasks to be performed using the LLM and, finally, define the LLM prompt that will perform the classification tasks. Please notice that all the software artifacts used to create this paper are available online in the following GitHub repository: https://github.com/grec-tcm/jnic2026.

4.1. CVE Database Creation

To assess the vulnerability of medical IT systems we start by building a database of related CVEs. We start the process by performing a product- and manufacturer-based approach in a CVE database (i.e., https://www.cve.org/CVERecord/). However, it turns out that this approach is not very scalable[1], as one needs to know which exact medical products and companies to look for in advance.

Alternatively, we decided to perform a semantical search on the CVE database using various keywords that are related to the medical IT systems presented in Section 3, namely[2]: Medical (318), Patient (267), Hospital (264), DICOM (102), PACS (93), EHR (27), EMR (21), FHIR (15), and HL7 (13). We also include the following terms that are related to the medical domain: Insulin (19), X-RAY (11), MRI (8), Implantable (5). In contrast, the term HIS (288), RIS (14), LIS (3), PIS (1), that are also related to the medical domain, are NOT included in the CVE database because after manual inspection the majority of results are not related to medical IT systems. Using such approach we obtain a total of 1163 CVE entries that are stored in texts files. However, these CVE entries can appear multiple times since they may appear in the results of multiple keywords searches. Hence, we sort them and remove the duplicates, ending up with 1097 unique CVE entries.

Next, we use a Python script to download the CVE information from the MITRE CVE database (https://cveawg.mitre.org/api/cve/) since it returns the CVE entries as JSON objects that can be later processed automatically. However, a comparative analysis of the returned JSON objects against external reference databases reveals significant inconsistencies in the data reported for identical CVEs. For example, for some CVE entries multiple CVSS versions are reported (i.e., 4.0, 3.1, 3.0 or 2.0), whereas for other CVE entries only one is reported. Hence, we decided to also download the CVE information from the NVD database (https://services.nvd.nist.gov/rest/json/cves/) and combine both databases to have as much information as possible for each CVE.

4.2. LLM Setup and Configuration Parameters

The first step to create a setup is to decide whether to run the experiments on the cloud using a commercial LLM, or locally using an open source LLM. The former has the benefit of allowing access to the latest LLM models, which provide the best features in terms of functionality and performance. However, using commercial LLMs increases the cost and limits the reproducibility of the experiments. Hence, we decide to run the LLM models locally using a workstation equipped with an AMD Ryzen 9 5950X processor, 32 GB of system memory, and an NVIDIA RTX 5070 GPU (Blackwell, 12 GBytes VRAM). This configuration enables high-throughput and low-latency execution of LLM models without additional cost or imposed rate limitations. In addition, having a controlled environment allows ensuring reproducibility of experiments that could be caused by external service variability.

Once using a local setup to run the experiments has been decided, the next step is to select which open source LLM model to use. There are many LLMs each having its own benefits and drawbacks in terms of capabilities, performance and hardware requirements [14]. After considering the requirements of the experiments and the constraints of our hardware platform, we have selected Gemma3 [15], an open-weight LLM developed by Google DeepMind. Thanks to the availability of a 12B-parameter version and its extended context capacity of up to 128k tokens, the model offers a practical balance

[1] Using such approach we retrieved 35 CVEs, which were also included to the CVE database.
[2] The number in parenthesis represents the number of CVE entries that each keyword search returned.

 doi:10.20944/preprints202603.0737.v1

between reasoning capability and computational cost, allowing to process large CVE entries while fitting within 12 GB of GPU VRAM, and allowing the concurrent execution of two workers while keeping inference fully on-GPU. This substantially reduces the time and the cost to perform the classification tasks, allowing to complete the tasks in approximately 160 minutes and consuming about 250 Watts of power.

To run the gemma3:12b LLM model using our hardware platform we use `ollama`, an open-source framework and platform designed to download, run, and manage LLMs locally. To ensure proper execution of the LLM in our experimental setup we perform some runtime optimizations of `ollama`. First, the maximum context length is fixed at 40,960 tokens (`OLLAMA_CONTEXT_LENGTH=40960`) to ensure that the LLM fits within the available GPU memory and precisely matches the prompt design without allocating unnecessary context capacity. Second, the key-value cache uses 4-bit quantization (`OLLAMA_KV_CACHE_TYPE="q4_0"`), which substantially reduces VRAM consumption by compressing attention cache storage. Third, flash attention is enabled (`OLLAMA_FLASH_ATTENTION=1`) to accelerate attention computations and increase memory efficiency, which is particularly beneficial for long-context inference. Finally, parallel execution is configured with two workers (`OLLAMA_NUM_PARALLEL=2`) to increase throughput while staying within hardware limits. All these parameters allow improving speed and stability of the execution, while staying within the GPU memory budget limitations. In addition to performance-related configuration parameters, we also set the sampling temperature to 0, enforcing deterministic behavior and minimizing variability across classification outputs, which ensures experimental stability and reproducibility.

4.3. LLM-Based Classification of CVEs

Following the preliminary database analysis, we leverage LLMs to perform a multi-dimensional classification of the collected CVEs. The exponential growth in reported vulnerabilities makes manual triaging unscalable, and traditional keyword-based filtering is inherently prone to high false-positive rates. For instance, a CVE description might contain medical-specific terminology (i.e., DICOM), but the actual vulnerability may reside in a general-purpose packet analyzer (e.g., Wireshark) rather than a clinical device. To address this limitation, the first classification task utilizes the LLM's contextual understanding to perform a binary assessment, effectively filtering the dataset to isolate vulnerabilities that genuinely impact the medical domain.

Once the true-positive medical CVEs are identified, the LLM performs a granular architectural classification to provide actionable context for medical IT systems. The model categorizes each verified CVE into specific medical IT system domains (i.e., data identity and management, departmental information systems, data generation systems, storage and archiving systems, and interoperability systems). At the same time, the LLM also classifies each CVE based on the specific technological component it affects (i.e., hardware, firmware, operating systems, software, or libraries). This dual-layered classification is highly relevant for medical environments, as it allows hospital IT teams to accurately pinpoint where a vulnerability resides within their infrastructure and assess its potential impact. Finally, to contextualize the threat landscape, the LLM maps each classified CVE to the MITRE ATT&CK framework. By categorizing vulnerabilities into specific adversarial tactics (reconnaissance, resource development, initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, command and control, exfiltration, and impact) the LLM translates static vulnerability descriptions into dynamic threat intelligence. This mapping is critical for securing medical IT systems, as it empowers defenders to understand exactly how a vulnerability might be weaponized as part of a broader intrusion campaign, ultimately guiding more effective and prioritized patching efforts based on realistic adversary behaviors.

4.4. LLM Prompt Development

After defining the platform and classification tasks, the next step is to design the prompt used to perform the classification. To ensure terminological consistency, all prompts are issued in English and aligned with CVE descriptions and the MITRE ATT&CK taxonomy. Each query follows a standardized

 doi:10.20944/preprints202603.0737.v1

template in which explicit classification instructions precede complete CVE records obtained from the NVD and MITRE databases. Template reuse minimizes prompt drift and promotes stable semantic interpretation across runs. The prompt has been iteratively refined to support consistent classification of vulnerabilities affecting healthcare environments. Healthcare relevance was operationalized using a strict *primary-purpose* criterion that distinguishes specialized clinical systems (e.g., EHR/EMR platforms, PACS, medical device firmware, and clinical middleware) from general-purpose software. A collateral-support exclusion rule prevents misclassification of tools that merely process medical data formats (e.g., DICOM or HL7). To reduce ambiguity, all classification dimensions are restricted to predefined options with scope notes and exclusions, including relevance, vendor and product identification, system category, affected component, and ATT&CK tactic mapping. MITRE ATT&CK tactic assignment follows a kill-chain priority rule that selects the earliest applicable attacker objective, promoting consistent tactical labeling when multiple exploitation paths exist. Finally, the prompt enforces a fixed input and output JSON schema, and requires the LLM to justify each decision using evidence from the CVE context, as well as to provide confidence scores, which enables auditing and reproducibility of the experiments results.

5. Results

This section presents the results of the CVE database analysis. The results are presented in two subsections. First, a preliminary analysis intended to understand the contents of the CVE database. Second, an LLM-enabled analysis of the CVE database intended to perform various classification tasks. Finally, we present a discussion of the obtained results.

5.1. Preliminary Database Analysis

After applying the filtering procedure described earlier, our CVE database contains a total of 1097 valid entries.

To analyze the database we start by plotting the number of CVEs grouped by the year that they have been first reported in the CVE database, as depicted in Figure 1. It is important to point out the rapid growth in CVEs being reported, going a total of 19 CVEs being reported between 2004 and 2010, to a total of 298 CVEs being reported in 2025 alone. As the number of CVEs reported per year increases, the pressure on cybersecurity analysts working in healthcare rises, as the surface attack grows and the time to patch the system increases, even without adding new systems.

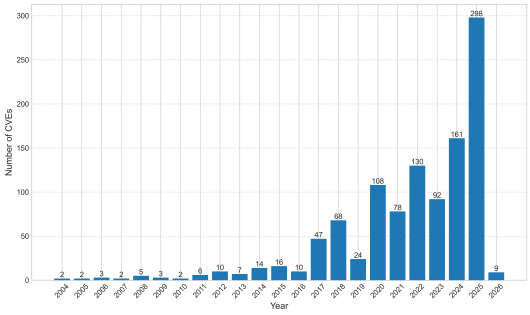

Figure 1. Number of CVE vulnerabilities that have been discovered and reported per year.

Next, we focus on the CVSS value of each CVE in the database. Figure 2 shows the CVE distribution according to their CVSS value (i.e., from 0 to 10) and CVSS category (i.e., low, medium, high or critical). Please note that the CVSS value reported for each CVE is the one from the latest version (e.g., CVSS 4.0, 3.1, 3.0 or 2.0) available in the database. As it can be observed, 17.5% (192) of CVEs belong to the critical category (i.e., CVSS value equal or greater than 9.0) and 30.9% (339) belong to the high category (i.e., CVSS value between 7.0 and 8.99). The fact that 48.4% of the CVEs analyzed belong to these categories puts even more pressure on cybersecurity analysts working in healthcare,

 doi:10.20944/preprints202603.0737.v1

since these threats need to be detected and patched effectively to minimize their potential impact on medical IT systems.

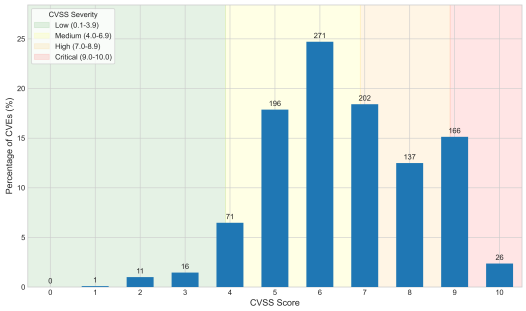

Figure 2. CVE distribution according to their CVSS value and category. Background colors (green, yellow, orange and red) represent CVSS severity (low, medium, high and critical).

5.2. LLM-Based Database Analysis

During preparation of the CVE database, we identified several entries that are not strictly related to medical IT systems. For example, `CVE-2021-21807` describes a vulnerability that affects a DICOM parsing component and was initially considered relevant to the medical sector because it contained a keyword used to construct the CVE database. However, upon closer inspection, we found that the vulnerability actually pertains to the Accusoft ImageGear DICOM component, part of a general-purpose image processing toolkit, and therefore should not be classified as relevant to the healthcare sector. Given the growing number of CVEs, automating this task is important, as it allows focusing specifically on the CVEs that are relevant to the medical sector.

Hence, the first step in our LLM-based database analysis is to classify which CVEs are related to medical IT systems and which are not. By running the prompt described in the previous section, we obtain the data depicted in Figure 3 and summarized in Table 1, which shows the number and percentage of CVEs that have been classified as medical-related and non-medical related. As it can be observed, the LLM has detected a total of 11.30% CVEs (124) that are not relevant to medical IT systems.

Table 1. Distribution of medical-related vs. non-medical related CVEs according to their CVSS value.

Value	Med (N)	Med (%)	Non-Med (N)	Non-Med (%)
0	0	0.0%	0	0.0%
1	1	0.1%	0	0.0%
2	8	0.7%	3	0.3%
3	13	1.2%	3	0.3%
4	58	5.3%	13	1.2%
5	167	15.2%	29	2.6%
6	244	22.2%	27	2.5%
7	178	16.2%	24	2.2%
8	129	11.8%	8	0.7%
9	149	13.6%	17	1.5%
10	26	2.4%	0	0.0%
Total	**973**	**88.7%**	**124**	**11.3%**

 doi:10.20944/preprints202603.0737.v1

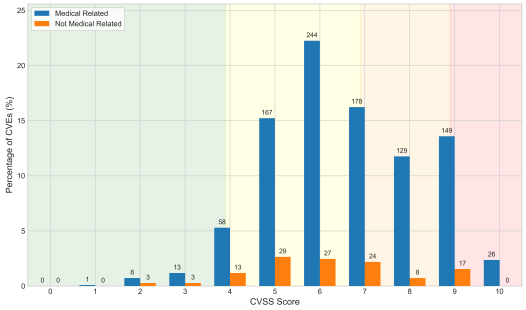

Figure 3. CVE distribution according to their CVSS value and category after filtering for medical relevance. Orange bars indicate CVEs that have been filtered out by the LLM.

To validate the performance of the LLM-based classification task, we conduct a manual classification by reviewing each CVE individually and applying the same criteria provided to the LLM in the prompt. This human-generated classification serves as the baseline for evaluating the model's performance in the initial binary task (e.g., healthcare-related vs. non-healthcare-related). The comparative results are summarized in Tables 2 and 3. As it can be observed, the model demonstrates strong predictive capabilities in classifying medical-related CVEs, achieving an overall accuracy of 92.00% and a robust F1-score of 95.60%. Coupled with a high recall (sensitivity) rate of 93.25%, the system successfully identified the vast majority of target CVEs, missing only a small fraction as false negatives (6.29%). Overall, the results validate that the LLM is a highly effective and trustworthy tool for automating the domain-specific triaging of security vulnerabilities in the medical sector.

Table 2. Classification Summary and Error Distribution.

Category	Count	Percentage
True Positives (TP)	954	86.96%
True Negatives (TN)	55	5.01%
False Positives (FP)	19	1.73%
False Negatives (FN)	69	6.29%
Total	1097	100%

Table 3. Detection Performance Metrics.

Metric	Value
Precision	98.05%
Recall (Sensitivity)	93.25%
F1-score	95.60%
Accuracy	92.00%

Next, we use the LLM prompt to classify the CVEs into the medical system categories presented in Section 3 (e.g., *Data Identity and Management*, *Departmental Information Systems*, *Data Generation Systems*, *Storage and Archiving Systems*, and *Interoperability Systems*). Using the LLM we obtain the results depicted in Figure 4, which presents the dispersion of CVSS base values across the medical system groups. The percentage distribution, shown in Table 4, reveals that vulnerabilities are not uniformly represented among categories. *Identity and Management Systems* and *Data Generation Systems* exhibit the highest density of CVE observations, indicating that these domains constitute the most frequently affected components within the dataset. This concentration is consistent with the functional characteristics of such systems, which commonly integrate authentication mechanisms, patient data processing pipelines, device interfaces, and protocol implementations, all well-known sources of software weaknesses. High and critical severity vulnerabilities (CVSS $\geq$ 7.0) appear across all system categories, demonstrating that impactful security failures are not confined to any single class of medical infrastructure. *Storage and Archiving Systems* show a noticeable clustering of elevated CVSS values, a

 doi:10.20944/preprints202603.0737.v1

finding of particular relevance given the persistent and high-value nature of the clinical data managed by these systems. From a risk perspective, vulnerabilities affecting storage components may therefore introduce disproportionate operational and regulatory consequences. Finally, *Interoperability Systems*, while associated with fewer CVE entries, display a broad severity range. This variability suggests that although vulnerabilities are less frequent in this category, their potential impact remains significant. Considering that interoperability layers mediate cross-system communication and trust relationships, even isolated high-severity weaknesses may enable cascading security effects.

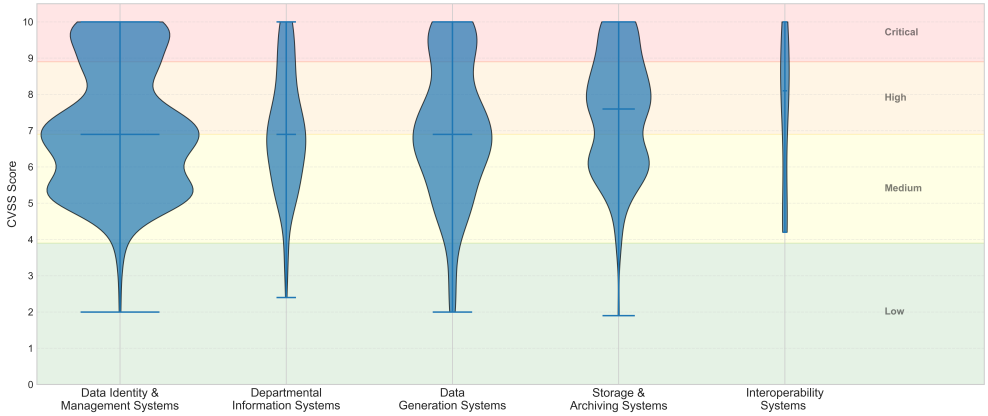

Figure 4. CVE distribution according to their CVSS value (0-10) and category (low, medium, high and critical), and the medical IT system category (Data Identity & Management, Departmental Information Systems, Data Generation Systems, Storage & Archiving Systems, and Interoperability Systems).

Table 4. Percentage of CVEs according to their CVSS category (low, medium, high and critical) and healthcare system category.

Category	Low (%)	Med (%)	High (%)	Crit (%)
Data Identity & Management Systems	1.4%	53.4%	27.3%	18.0%
Departmental Information Systems	3.7%	52.3%	31.2%	12.8%
Data Generation Systems	4.5%	45.9%	29.1%	20.5%
Storage & Archiving Systems	1.1%	37.8%	43.3%	17.8%
Interoperability Systems	0.0%	33.3%	45.8%	20.8%

Next, we use the LLM to classify the CVEs according to the affected component types (e.g., *Hardware*, *Firmware*, *Operating System*, *Software*, and *Library*). The LLM prompt returns the data presented in Figure 5 and summarized in Table 5. While it is interesting to observe that the LLM is able to complete the task, the results reveal a limitation in classifying the CVEs into their appropriate categories. Despite explicit instructions, the model did not reliably differentiate the categories *Operating Systems* and *Libraries* from the broader *Software* category, effectively merging multiple technical layers.

Table 5. Percentage of CVEs according to their category (low, medium, high and critical) and the affected component.

Component	Low (%)	Med (%)	High (%)	Crit (%)
Hardware	0.0%	50.0%	30.0%	20.0%
Firmware	3.9%	48.5%	27.2%	20.4%
Operating System	-	-	-	-
Software	2.1%	48.1%	32.1%	17.7%
Library	-	-	-	-

 doi:10.20944/preprints202603.0737.v1

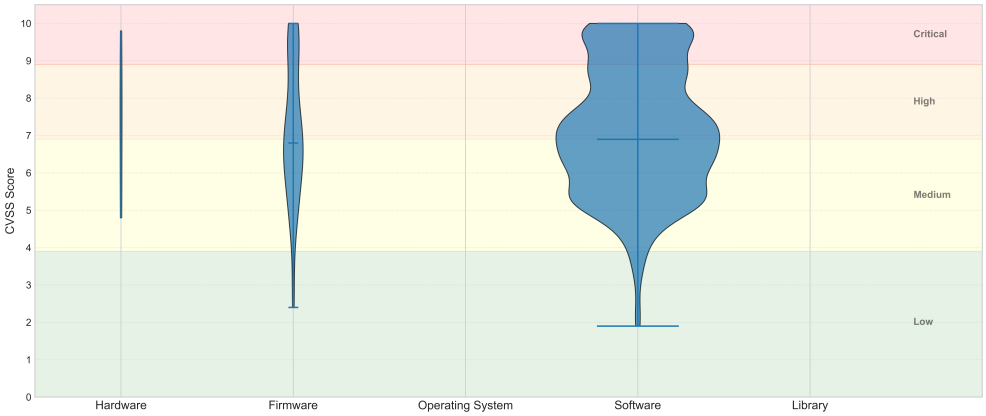

Figure 5. CVE distribution according to their CVSS value (0-10) and category (low, medium, high and critical), and the affected component (i.e., *Hardware*, *Firmware*, *Operating System*, *Software*, and *Library*).

Finally, we use the LLM to classify the CVEs within the MITRE ATT&CK framework to complement the system-centric perspective. Figure 6 illustrates the results, indicating that vulnerabilities are predominantly associated with early and mid-stage adversarial activities, particularly *Initial Access*, *Execution*, *Credential Access*, and *Discovery*. These phases exhibit both a high concentration of observations and a wide severity spread, including numerous high and critical CVSS values. This pattern aligns with the practical objectives of attackers, as weaknesses enabling entry, code execution, or credential compromise provide foundational capabilities for subsequent operations. Later-stage phases such as *Exfiltration and Impact* contain fewer observations but still include high-severity vulnerabilities. Although less frequent, weaknesses mapped to these stages may correspond to incidents with direct clinical or organizational consequences, including data loss, service disruption, or system manipulation. Overall, the ATT&CK-aligned analysis reinforces that vulnerability severity must be interpreted in operational context. The security significance of a CVE is not solely determined by its base value but also by the attacker behaviors it may facilitate. This adversarial mapping therefore provides a structured lens for understanding how technical weaknesses may translate into realistic threat scenarios within medical IT environments.

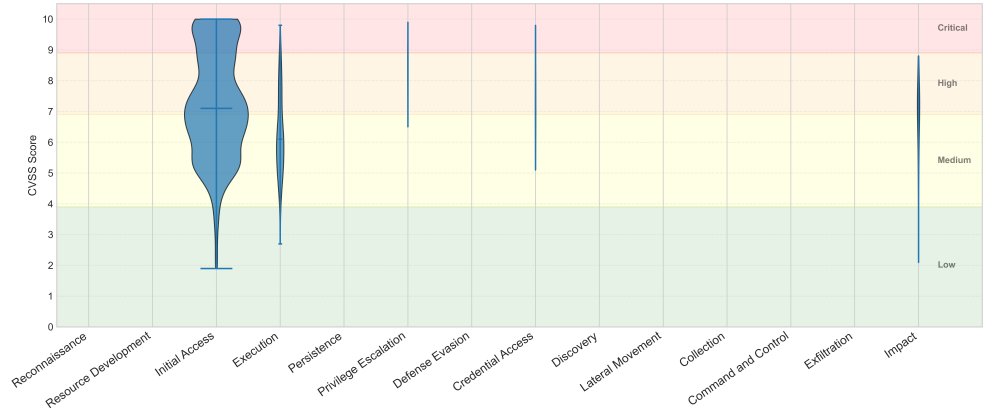

Figure 6. CVE distribution according to their CVSS value (0-10) and category (low, medium, high and critical), and the MITRE ATT&CK framework.

 doi:10.20944/preprints202603.0737.v1

Table 6. Percentage of CVEs according to their CVSS category (low, medium, high and critical) and the MITRE ATT&CK framework.

ATT&CK Phase	Low (%)	Med (%)	High (%)	Crit (%)
Reconnaissance	-	-	-	-
Resource Development	-	-	-	-
Initial Access	2.3%	46.5%	30.6%	20.6%
Execution	2.1%	67.7%	29.2%	1.0%
Persistence	-	-	-	-
Privilege Escalation	0.0%	12.5%	62.5%	25.0%
Defense Evasion	-	-	-	-
Credential Access	0.0%	33.3%	55.6%	11.1%
Discovery	-	-	-	-
Lateral Movement	-	-	-	-
Collection	-	-	-	-
Command and Control	-	-	-	-
Exfiltration	-	-	-	-
Impact	3.3%	46.7%	50.0%	0.0%

5.3. Discussion

The presented results demonstrate the potential of using LLMs to automatically classify CVEs relevant to the medical domain, while also extracting supplementary risk-assessment data absent from the original CVE entries. However, notable limitations remain. Even with the model's temperature parameter set to 0, the LLM occasionally hallucinates, yielding false positives and negatives. For instance, the model might incorrectly flag a CVE due to the presence of medical keywords, failing to recognize that the affected software is rarely used in clinical settings (a false positive). Conversely, it might overlook a vulnerability in a generic library that serves as a critical dependency for widely used medical software (a false negative). While these errors can be mitigated by engineering prompts with deep contextual knowledge, domain-specific vocabulary, and strict few-shot examples, overly complex prompting undermines the initial scalability and general utility of off-the-shelf LLMs. Furthermore, although we instructed the LLM to generate both a classification rationale and a confidence score, our retrospective analysis showed that while the reasoning was accurate and highly valuable for validation, the confidence scores were largely unreliable. Future research should evaluate how different LLM architectures and parameter sizes affect performance on this specific task. Additionally, exploring ensemble approaches—where outputs from multiple LLMs are aggregated—could significantly improve classification accuracy and reliability compared to standalone models.

6. Conclusions and Future Work

This article demonstrates the viability of utilizing LLM models to automate the triage and multidimensional classification of CVEs within the medical IT domain. Our framework successfully filtered healthcare-relevant vulnerabilities with an accuracy of 92.00% and a 95.60% F1-score, as well as effectively mapping them to specific functional medical systems and MITRE ATT&CK adversarial phases. This automated augmentation provides critical operational context often missing from standard CVE records, revealing that data identity systems and data generation equipment are frequent targets, predominantly during the initial access and execution stages of an attack. However, our analysis also highlights notable limitations in off-the-shelf LLM reasoning, including difficulties in granular component-level categorization (such as distinguishing between operating systems and libraries), occasional contextual hallucinations, and the unreliability of the model's self-reported confidence scores.

To address these limitations and further advance automated vulnerability assessment, future work will focus on comparative evaluations of diverse LLM architectures and parameter scales to determine the optimal configurations for domain-specific categorization. We also plan to explore ensemble-based classification methodologies, aggregating outputs from multiple LLM models to

 doi:10.20944/preprints202603.0737.v1

mitigate individual hallucinations and enhance overall diagnostic reliability. Additionally, refining context-aware prompting strategies to balance strict domain-specific vocabulary with the scalability of baseline models remains a primary objective. Finally, extending our framework to include deeper software dependency-chain analysis will help capture complex, indirect vulnerabilities, ultimately ensuring a more robust and comprehensive security posture for critical healthcare infrastructures.

Acknowledgments: This project is carried out within the framework of the Recovery, Transformation, and Resilience Plan, funded by the EU Next Generation funds, under the auspices of the INCIBE cybersecurity chair named CARISMATICA.

Referenecs

1. Walkowski, Michał et al.. Vulnerability Management Models Using a Common Vulnerability Scoring System. *11*, 8735. https://doi.org/10.3390/app11188735.
2. Boateng, Gordon Owusu et al.. A Survey on Large Language Models for Communication, Network, and Service Management: Application Insights, Challenges, and Future Directions. *IEEE Communications Surveys & Tutorials* **2026**, *28*, 527–566. https://doi.org/10.1109/COMST.2025.3564333.
3. Marchiori, Francesco et al.. Can LLMs Classify CVEs? Investigating LLMs Capabilities in Computing CVSS Vectors. In Proceedings of the 2025 IEEE Symposium on Computers and Communications (ISCC), 2025, pp. 1–6. https://doi.org/10.1109/ISCC65549.2025.11326354.
4. Miranda, Lucas et al.. Learning CNA-Oriented CVSS Scores. In Proceedings of the 2024 IEEE 13th International Conference on Cloud Networking (CloudNet), 2024, pp. 1–5. https://doi.org/10.1109/CloudNet62863.2024.10815736.
5. Mirtaheri, Seyedeh Leili et al.. Knowledge-Driven Large Language Models for Automating CVSS Score Prediction. In Proceedings of the Proceedings of the 2025 Workshop on Research on Offensive and Defensive Techniques in the Context of Man At The End (MATE) Attacks, New York, NY, USA, 2025; CheckMATE '25, p. 20–28. https://doi.org/10.1145/3733817.3762699.
6. Jafarikhah Sima, et al.. From Description to Score: Can LLMs Quantify Vulnerabilities?, 2026, [arXiv:cs.CR/2512.06781].
7. Tuset-Peiró, Pere et al.. Assessing cybersecurity of Internet-facing medical IT systems in Germany & Spain using OSINT tools. In Proceedings of the Actas de las X Jornadas Nacionales de Investigación en Ciberseguridad (JNIC), Zaragoza, Spain, 2025; pp. 190–197.
8. HL7 International. HL7 Version 2 Product Suite. https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185, 2023. Accessed: 2026-02.
9. *ISO/HL7 10781:2015*; Health Informatics — Electronic Health Record System Functional Model (EHR-S FM). International Organization for Standardization: 2015.
10. *ISO 12967:2012*; Health Informatics — Service Architecture (HISA). International Organization for Standardization: 2012.
11. DICOM Standards Committee. Digital Imaging and Communications in Medicine (DICOM) Standard. https://www.dicomstandard.org, 2023. Accessed: 2026-02.
12. Shortliffe, E.H.; Cimino, J.J. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, 4th ed.; Springer, 2014.
13. Huang, H.K. *PACS and Imaging Informatics: Basic Principles and Applications*, 2nd ed.; Wiley-Blackwell, 2010.
14. Naveed, H.e.a. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*. https://doi.org/10.1145/3744746.
15. Gemma Team et al.. Gemma 3 Technical Report, 2025, [arXiv:cs.CL/2503.19786].