MedAIScout: Automated Retrieval of Known Machine Learning Vulnerabilities in Medical Applications

Athish Pranav Dharmalingam
Department of Computer Science
Indian Institute of Technology Madras
Chennai, India
athishanna@gmail.com

Gargi Mitra
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
gargimiitm@gmail.com

Abstract

Machine learning (ML)-enabled medical devices are transforming the healthcare industry but are vulnerable to adversarial attacks that can compromise their safety. Current red teaming efforts often overlook these ML-specific threats, leaving devices exposed. To address this, we present MedAIScout, a semi-automated tool designed to retrieve information on known ML vulnerabilities relevant to ML-enabled medical devices. Through case studies on five FDA-approved ML-enabled devices, we demonstrate that MedAIScout effectively identifies relevant vulnerabilities in four of them, significantly aiding red teaming efforts.

1 Introduction

Machine learning (ML)-enabled medical devices are revolutionizing healthcare by offering high-precision diagnostics, personalized treatment plans, and even real-time surgical support [23]. At present, more than 900 such devices are registered with the US Food and Drug Administration (FDA). However, their reliance on complex and often unexplainable models exposes them to a range of attacks [8, 15] that might potentially target the ML algorithms and lead them to misdiagnose a patient's condition or suggest wrong treatment plans, putting lives at risk. Existing works [2, 4, 13, 1] demonstrate that poisoning training data or introducing small perturbations during prediction can drastically alter the outputs of ML models, which could remain undetected until it is too late. Most of the state-of-the-art red teaming exercises treat ML-enabled devices as equivalent to traditional software-driven systems. A prior work [18] attempts to automate the identification of security risks in different assets of a healthcare information infrastructure but does not consider attacks on the ML techniques used in these assets. These approaches overlook attack vectors specific to the ML techniques, leaving critical vulnerabilities unaddressed. To mitigate these risks, the Food and Drug Administration (FDA) of the USA [21] recommends addressing the security concerns of the AI [1] techniques used in medical devices, which can be achieved through rigorous AI red-teaming exercises [22].

AI red teams must simulate attacks that manipulate the training or testing data to compromise the integrity, reliability, and usability of an ML model [22]. For this, the red teams must first acquire a comprehensive list of different adversarial ML techniques and understand how they might affect the device under assessment. Existing works have leveraged the MITRE ATLAS knowledge base [31] for this, which recommends using the following sources – (i) victim's publicly available research materials and their websites, to understand how ML is incorporated into the system and other technical details of the ML-enabled products; and, (ii) journals and conference proceedings, preprint repositories, technical blogs and the victim's public research articles that might contain information

[1] Despite their technical differences, the terms Artificial Intelligence (AI) and Machine Learning (ML) have been used interchangeably in this paper.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

about the vulnerabilities of the common ML models used in the device. However, with the vast body of literature and rapid increase in published attacks targeting ML models, manually browsing through these materials is not only time-consuming but also highly prone to errors. Therefore, an automated process for discovering and filtering information is essential to ensure the timely and accurate identification of threats. This would enable red teams to stay updated on emerging attacks and reduce the time required to assess a device.

This paper introduces MedAIScout, a semi-automated tool designed to retrieve information about known machine learning (ML) vulnerabilities that could impact AI-powered medical devices, thereby supporting red-teaming efforts. Given the description of an ML-enabled medical device, MedAIScout first applies Natural Language Processing (NLP) techniques to identify key terms that describe the device's functionality, the type of ML model used, and the nature of the data it processes. Using these extracted terms, MedAIScout constructs tailored search queries to retrieve peer-reviewed research articles that document attacks on the ML model employed by the device. Additionally, MedAIScout distinguishes between training-time and inference-time attacks, providing context and explanations for the relevance of each retrieved article to the specific device. Over the device's lifecycle, a red team can periodically use MedAIScout to monitor for new ML vulnerabilities pertinent to the device. To our knowledge, MedAIScout is the first automated tool specifically designed to retrieve information on known ML vulnerabilities within the context of medical applications.

Designing MedAIScout presents two main challenges: **(i)** The lack of a standardized format for manufacturers to disclose device descriptions. As a result, these documents are often highly unstructured, causing the NLP technique to produce incomplete or fragmented phrases that hinder the creation of effective search queries. To address this, MedAIScout is integrated with publicly available large language models (LLMs), such as ChatGPT, to filter out irrelevant content and repair broken phrases; and **(ii)** Manufacturers provide varying levels of detail about the ML techniques used in their devices. Some specify the exact ML architecture, while others use broad terms like 'machine learning' or 'artificial intelligence'. When only generic terms are available, MedAIScout instead leverages information about the device's functionality and the type of data it processes to retrieve relevant research articles.

We evaluate MedAIScout on 5 different randomly chosen FDA-approved ML-enabled devices - (i) Clarius Ultrasound Scanner, (ii) Auto Segmentation (a radiology device), (iii) Quantib Prostrate, (iv) Volpara Imaging Software, and (v) Caption Interpretation Automated Ejection Fraction Software. Our results show that for each of the 5 devices, MedAIScout was able to retrieve at least one recent inference-time attack paper and/or at least one recent training-time attack paper. As extra information, MedAIScout also retrieved some useful defense papers in both categories for 4 of the 5 devices assessed.

Contributions. Our contributions are summarized as follows.

1. We propose MedAIScout, a semi-automated tool that retrieves information on known ML vulnerabilities and attacks that could compromise a specific ML-enabled medical device;
2. We develop techniques to address the lack of standardized device description formats and varying levels of detail in manufacturers' disclosures, ensuring that relevant articles are retrieved for the ML-enabled device under assessment; and,
3. We conduct case studies on real ML-enabled devices, demonstrating that MedAIScout can retrieve pertinent articles on ML attacks within a few seconds, significantly reducing the red team's effort.

2 MedAIScout

We propose MedAIScout, an automated tool for retrieving information on known ML vulnerabilities for a given ML-enabled medical device. Figure 1 shows the workflow of MedAIScout. In this section, we will first discuss the rationale behind the workflow, followed by our design choices at each step.

2.1 Rationale behind MedAIScout workflow

As a preliminary approach, we initially attempted to leverage the capabilities of publicly accessible or open-source Large Language Models (LLMs) such as ChatGPT [32], Llama [30] and Gemini [26],

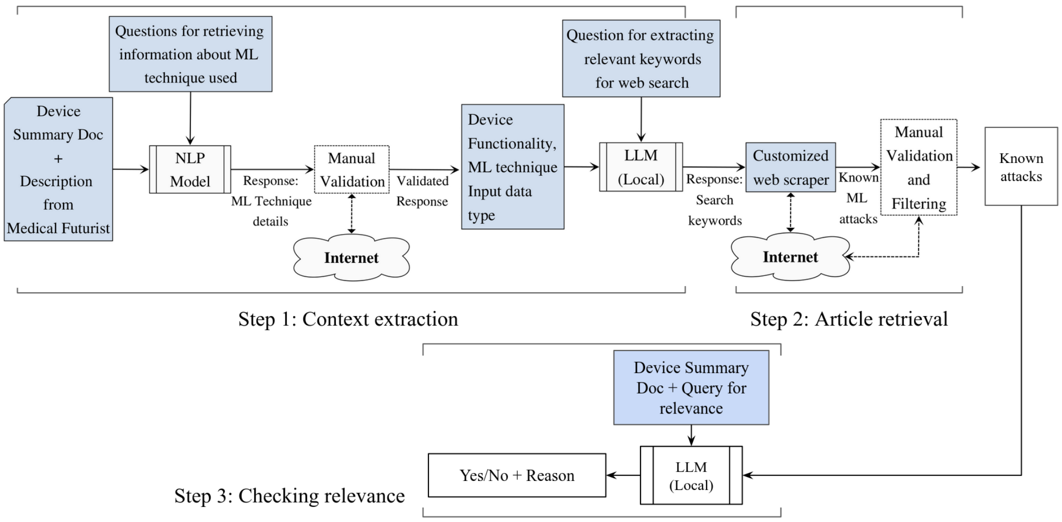
Questions for retrieving information about ML technique used
Question for extracting relevant keywords for web search
Device Summary Doc + Description from Medical Futurist
NLP Model
Response: ML Technique details
Manual Validation
Validated Response
Internet
Device Functionality, ML technique Input data type
LLM (Local)
Response: Search keywords
Customized web scraper
Known ML attacks
Manual Validation and Filtering
Internet
Known attacks
Step 1: Context extraction
Step 2: Article retrieval
Device Summary Doc + Query for relevance
Yes/No + Reason
LLM (Local)
Step 3: Checking relevance

Figure 1: MedAIScout Workflow

to directly retrieve relevant articles from the web based on the device description. However, this single-step solution presented several limitations.

- While ChatGPT was effective in parsing unstructured text from the device documentation, it encountered issues when retrieving the final list of articles. Specifically, it produced hallucinations (false results) and often generated inconsistent output across different runs.
- Llama does not provide the feature of parsing text from input files. Furthermore, the local models are not updated with the latest articles and often hallucinate.
- The basic version of Gemini does not support text parsing from input files, and the advanced version has a limit on the maximum volume of data that can be uploaded.

Given the aforementioned limitations and the need for usability in red-teaming efforts, we designed MedAIScout to be user-friendly and free from false information. Therefore, we broke down the task of identifying relevant information from the device description and using it to retrieve relevant articles on ML attacks into smaller, manageable subtasks. Each subtask is tractable and can be enhanced individually in the future, if necessary.

2.2 Step 1: Context Extraction

In this step, MedAIScout extracts the device functionality, ML technique used, and input data types from the device descriptions provided by the manufacturer, in line with the recommendations from MITRE [31]. We use two sources of information - publicly available summary documents submitted by the manufacturer to the FDA along with their approval request, and a more streamlined version of this information hosted in a tabular fashion on a public platform by Medical Futurist [9].

To parse the input text and extract relevant information we leverage DistilBERT base uncased distilled SQuAD [7], a lightweight and faster version of the BERT NLP model, which is well-known for its context interpretation capabilities. We ask the following contextual questions to the BERT model – “What is the device functionality?", “What is the machine learning technique used?", and “What is the input data type for this device?". Note that, we chose an exhaustive set of questions that cover multiple aspects of the ML technique. Out of the two versions of the NLP model, we select the one trained with the SQUAD1 dataset[3], as the version trained on the SQUAD2 dataset[6] is trained to answer even unanswerable questions, and might generate false and irrelevant output if it does not find the exact answers.

However, extracting the necessary information from these documents presents several challenges. First, the summary documents are typically lengthy (often exceeding 10 pages), with technical details sparsely distributed throughout. As a result, NLP models struggle to generate accurate answers when processing the entire document as input. To mitigate this issue, we provide each paragraph in a summary document as input at a time, applying the question-answering process to each paragraph individually. The results are then aggregated, with duplicate answers removed, to ensure completeness and improve the overall accuracy of the extracted information.

Second, the summary documents rarely include explicit detailed technical details about the device's ML algorithms, and Medical Futurist published data for only 10% of the total number of approved devices, leaving a gap in up-to-date information. Additionally, the documents provided by manufacturers are often in PDF format with inconsistent fonts and formatting. This inconsistency causes parsing difficulties, resulting in incomplete and incomprehensible words and phrases. Moreover, the emphasis on functional descriptions of the device, rather than its technical aspects, further complicates the extraction of relevant data. To address this, we recommend the MedAIScout users manually inspect the output and eliminate the irrelevant words (e.g., standalone articles, malformed words) that are obvious. The remaining answers generated by the NLP model are passed through a large language model (LLM) to identify keywords and phrases relevant to ML algorithms. This step helps eliminate unnecessary text and correct broken words or phrases. We utilize Llama3[20] for this task, which, being open-source and locally hosted, ensures data security—a critical concern in the medical domain, and offers cost-efficiency.

2.3 Step 2: Retrieval of peer-reviewed articles on ML attacks

MedAIScout uses the device details identified in Step 1 to build web search queries in this step. It starts with the ML technique-related words and phrases identified in Step 1, and searches for recent attack papers on Google Scholar [27]. In cases where the device documentation does not provide sufficient details about the ML technique, this query might not return relevant papers. Therefore, MedAIScout also builds a query for retrieving attack papers on devices that have the same medical functionality and process the same type of data as the device under assessment. The user can configure the number of top search results MedAIScout should return.

Once MedAIScout returns the search results, the user is recommended to manually inspect the titles of articles to gauge their relevance and eliminate the obviously irrelevant articles (e.g., medical articles with no mention of security). The abstracts of each of the remaining articles are again passed to a local LLM (Llama3[20]) with a prompt asking it to specify if the article describes a training-time attack or an inference-time attack.

2.4 Step 3: Checking the relevance of retrieved articles

We again leverage a local LLM (Llama3[20]) to verify the relevance of each article retrieved in the previous step to the device under assessment. We pass the device summary, the abstract of the article, and a query asking for comments on the relevance of the article as inputs to the LLM. The LLM output is either yes or no, along with a short justification for the decision.

3 Evaluation

We evaluated MedAIScout on the following 5 ML-enabled medical devices.

1. Auto Segmentation by GE Medical Systems[33] - It is a software that processes CT scan images to detect organs at risk.
2. Clarius Ultrasound Scanner by Clarius Mobile Health Corp.[25] - It is an ultrasound imaging system.
3. Quantib Prostate by Quantib BV[24] - It is an image post-processing software for visualizing Prostate MRI images.
4. Volpara Imaging Software by Volpara Health Technologies Limited[29] - This software processes raw data from digital breast X-ray systems to estimate breast tissue density, which aids healthcare professionals in the assessment of breast composition.
5. Caption Interpretation Automated Ejection Fraction Software by Caption Health, Inc.[28] - This is a radiological quantitative imaging software that assists clinicians in cardiac evaluation.

Table 1: MedAIScout evaluation results on 5 different ML-enabled FDA-approved medical devices. ^{T} indicates articles on training-time attacks, and ^{I} indicates articles on inference-time attacks.

Device	Context extraction		Articles retrieved	Comments on relevance provided by MedAIScout
	Device functionality and data type	ML technique used		
Auto Segmentation	✓	✓	Aleksander Mądry et al.[5]^I	The adversarial attack described in the paper is theoretically applicable to the Auto Segmentation device, given its reliance on deep learning algorithms. The extent to which it is vulnerable would depend on the specific implementation details of the deep learning models, the presence of any security measures, and the effectiveness of the practitioner review process in identifying and correcting any errors caused by adversarial inputs.
			S Al-Zu'bi et al.[17]^T	The attack described in the new paper is not directly applicable to the Auto Segmentation device mentioned earlier. The new paper proposes a novel system for 3D image segmentation, which focuses on improving efficiency, accuracy, and security using distributed computing and secure transmission methods. The Auto Segmentation device relies on 2D deep learning-based segmentation, so the specifics of the 3D segmentation approach using HMMs do not directly relate to it. However, understanding the security measures and techniques in the context of distributed processing could be indirectly beneficial if the Auto Segmentation system were to evolve towards supporting 3D images or distributed architectures.
Clarius Ultrasound Scanner	✓	×	Angona Biswas et al.[19]^I	The adversarial attack strategies discussed in the paper are potentially applicable to the Clarius Ultrasound Scanner, particularly if it or any connected systems utilize DNNs or ML algorithms for image analysis or post-processing. The use of wireless communication for transmitting images adds further risk, as it creates additional points of vulnerability where adversarial manipulations could be introduced. The robustness of the Clarius Ultrasound Scanner against such attacks would depend on the specific security measures implemented in both the device and any connected ML-based analysis systems.
			Munachiso Nwadike et al.[10]^T	The backdoor attack described in the paper is potentially applicable to the Clarius Ultrasound Scanner, especially if it or its connected systems use deep learning models that could be trained or retrained using data that might be vulnerable to tampering. The primary concern is that any ML-based diagnostic or image analysis system connected to the Clarius device could be compromised if attackers gain access to the training data. As such, robust security measures, including secure data handling, training protocols, and the use of explainability techniques to detect unusual model behavior, are critical to mitigate the risk of backdoor attacks in medical imaging systems like the Clarius Ultrasound Scanner.
Quantib Prostate	✓	×	Alexander Ziller et al.[16]^I	The privacy risks and attacks discussed in the paper are potentially applicable to the Quantib Prostate device, particularly if it involves the use of federated learning or other collaborative ML techniques for model training. If the system uses federated learning without adequate privacy protections, such as differential privacy, it could be vulnerable to privacy attacks, including model inversion. To ensure the secure deployment of the Quantib Prostate in clinical settings, incorporating privacy-enhancing technologies is crucial to protect patient data while maintaining high segmentation and diagnostic performance.
			Laura Daza et al.[12]^T	The adversarial robustness framework and attack strategies discussed in the paper are highly relevant to the Quantib Prostate device. Since the device relies on deep learning for segmentation of prostate MRI images, its models could be vulnerable to adversarial attacks. Using the new benchmark proposed in the paper to evaluate the robustness of Quantib Prostate's models against a variety of adversarial perturbations could help improve their resilience and ensure accurate, reliable performance in clinical settings. Additionally, the novel ROG architecture could be explored as a potential solution for enhancing robustness across different medical segmentation tasks.
Volpara Imaging Software	✓	×	Georgios Kaissis et al.[14]^I	The paper is highly relevant to the Volpara Imaging Software, particularly in the context of implementing privacy-preserving techniques for handling sensitive medical imaging data. Integrating similar privacy and security measures could significantly enhance the confidentiality and integrity of data processed by Volpara Imaging Software, aligning with best practices for handling medical data securely.
			Ibrahim Yilmaz et al.[11]^T	The paper is highly relevant to the Volpara Imaging Software. While the specific focus is on mammographic image classifiers, the principles of adversarial attacks and their potential impact on medical imaging systems apply broadly. Ensuring that the Volpara Imaging Software is resilient to adversarial perturbations is essential for maintaining the accuracy and reliability of its volumetric assessments and density measurements. The paper highlights the importance of understanding and mitigating the risks associated with adversarial attacks in the context of medical imaging.
Caption	✓	×	No relevant papers retrieved	-

In the case of all the aforementioned devices, mispredictions can potentially result in misdiagnosis.

Table 1 shows the evaluation results. While MedAIScout retrieved multiple articles for many of the devices, in this table we present only one training-time attack article and one inference-time attack article in the interest of space. The following are the key takeaways from the evaluation.

1. For all the devices, MedAIScout was able to infer the device details from the summary documents.
2. However, MedAIScout could not retrieve the ML technique used by 4 of the devices from their publicly available descriptions used in our experiments. Manual inspection showed that the device descriptions either did not mention the ML technique (for instance, in the case of Caption) or only mentioned some generic terms such as 'machine learning' or 'artificial intelligence' (for instance, in the case of Quantib Prostrate). Under such circumstances, MedAIScout proceeded to retrieve papers that demonstrated attacks on ML-based techniques used for performing a task similar to the one performed by the device under assessment.
3. Most of the time the device description was sufficient to identify relevant articles on attacks based on their applicability in ML-enabled devices with similar functionalities. However, in the case of Caption, MedAIScout did not find any paper that demonstrated an attack on an ML technique that performs automated estimation of left ventricular ejection fraction. This is not a major limitation of MedAIScout though, as we were unable to retrieve attack papers relevant to this topic even with extensive manual search.
4. Upon appropriate prompts containing device description and the attack paper abstract, MedAIScout identified the key points for relevance with reasoning.

Implementation and evaluation environment. MedAIScout has been developed in Python3. For evaluation, it was executed on a local system with the following configurations – CPU: 11th Gen Intel Core i7-11390H, GPU: Intel Xe Graphics (TGL GT2), Ram: 16GB DDR4, OS: Fedora 40(Workstation), Kernel: Linux 6.10.9-200.fc40.x86_64, Disk: 512GB SSD. Note that MedAIScout has also been tested on Ubuntu 22.04. Evaluation on each device took $\approx 10 - 20$ minutes. While the actual process required only a few minutes, we added delays to the web scraping script to avoid overwhelming the network and server.

4 Discussion

While these are not fundamental limitations of our work, the efficiency of MedAIScout can be further enhanced by addressing the following concerns.

Limited access to information. While we only used publicly available sources for obtaining device descriptions in our evaluation, the red teams hired by the medical device manufacturers will have access to more detailed and structured documentation. We anticipate that under such circumstances MedAIScout will be able to generate more useful and accurate results.

Need for manual intervention. MedAIScout is not yet completely automated, and manual interventions are still required to verify some of the intermediate results. However, the effort and time involved in doing so is negligible (a few minutes' worth of work) as compared to the time and effort saved by MedAIScout (a few hours' worth of manual work for each device).

Need for extensive evaluation and comparative study. While MedAIScout successfully retrieved relevant articles for 4 devices, it would be interesting to observe the results of integrating different NLP models and LLMs. Furthermore, we plan to conduct extensive evaluations of MedAIScout using a diverse range of devices encompassing various physiological domains and tasks. This will give us important insights such as the class of devices that are the most vulnerable, the information that is crucial for retrieving the most relevant attack papers, etc.

5 Conclusion

In this paper, we present MedAIScout, a semi-automated tool for assisting AI red teams in performing a security risk assessment of ML-enabled medical devices. MedAIScout leverages publicly accessible sources of information, which it processes using NLP techniques and LLMs to retrieve the latest information on known ML vulnerabilities that are relevant to the device being assessed. We evaluated MedAIScout on 5 real ML-enabled medical devices. The results demonstrated MedAIScout's capability in retrieving useful information for the red teams for 4 out of 5 devices.

In the future, we will evaluate MedAIScout on a wider variety of medical devices, and explore its suitability in other application areas that use ML-enabled devices.

References

[1] Michael Backes et al. "Membership Privacy in MicroRNA-Based Studies". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: Association for Computing Machinery, 2016, pp. 319–330. ISBN: 9781450341394. DOI: 10.1145/2976749.2978355. URL: https://doi.org/10.1145/2976749.2978355.

[2] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. "Adversarial Machine Learning at Scale". In: *CoRR* abs/1611.01236 (2016). arXiv: 1611.01236. URL: http://arxiv.org/abs/1611.01236.

[3] Pranav Rajpurkar et al. "SQuAD: 100, 000+ Questions for Machine Comprehension of Text". In: *CoRR* abs/1606.05250 (2016). arXiv: 1606.05250. URL: http://arxiv.org/abs/1606.05250.

[4] Xinyun Chen et al. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning". In: *CoRR* abs/1712.05526 (2017). arXiv: 1712.05526. URL: http://arxiv.org/abs/1712.05526.

[5] Aleksander Mądry et al. "Towards deep learning models resistant to adversarial attacks". In: *stat* 1050.9 (2017).

[6] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *CoRR* abs/1806.03822 (2018). arXiv: 1806.03822. URL: http://arxiv.org/abs/1806.03822.

[7] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *NeurIPS EMC*^{2}$Workshop$. 2019.

[8] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[9] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database". In: *npj Digital Medicine* 3.1 (Sept. 2020), p. 118. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00324-0. URL: https://doi.org/10.1038/s41746-020-00324-0.

[10] Munachiso Nwadike et al. "Explainability matters: Backdoor attacks on medical imaging". In: *arXiv preprint arXiv:2101.00008* (2020).

[11] Ibrahim Yilmaz. "Practical fast gradient sign attack against mammographic image classifier". In: *arXiv preprint arXiv:2001.09610* (2020).

[12] Laura Daza, Juan C Pérez, and Pablo Arbeláez. "Towards robust general medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer. 2021, pp. 3–13.

[13] Matthew Jagielski et al. "Subpopulation Data Poisoning Attacks". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. CCS '21. Virtual Event, Republic of Korea: Association for Computing Machinery, 2021, pp. 3104–3122. ISBN: 9781450384544. DOI: 10.1145/3460120.3485368. URL: https://doi.org/10.1145/3460120.3485368.

[14] Georgios Kaissis et al. "End-to-end privacy preserving deep learning on multi-institutional medical imaging". In: *Nature Machine Intelligence* 3.6 (2021), pp. 473–484.

[15] Adnan Qayyum et al. "Secure and Robust Machine Learning for Healthcare: A Survey". In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 156–180. DOI: 10.1109/RBME.2020.3013489.

[16] Alexander Ziller et al. "Differentially private federated deep learning for multi-site medical image segmentation". In: *arXiv preprint arXiv:2107.02586* (2021).

[17] Shadi Al-Zu'bi et al. "Efficient 3D medical image segmentation algorithm over a secured multimedia network". In: *Multimedia Tools and Applications* 80 (2021), pp. 16887–16905.

[18] Stefano Silvestri et al. "A Machine Learning Approach for the NLP-Based Analysis of Cyber Threats and Vulnerabilities of the Healthcare Ecosystem". In: *Sensors* 23.2 (2023). ISSN: 1424-8220. DOI: 10.3390/s23020651. URL: https://www.mdpi.com/1424-8220/23/2/651.

[19] Angona Biswas et al. "Securing the Diagnosis of Medical Imaging: An In-depth Analysis of AI-Resistant Attacks". In: *arXiv preprint arXiv:2408.00348* (2024).

[20] Abhimanyu Dubey et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[21] US Food, Drug Administration, et al. *Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and OCP are Working Together*. 2024.

[22] Anna Raney et al. "An AI red team playbook". In: *Assurance and Security for AI-enabled Systems*. Vol. 13054. SPIE. 2024, pp. 71–97.

[23] U.S. FDA. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (visited on 10/31/2024).

[24] Quantib BV. *Quantib Prostate*. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K202501 (visited on 10/31/2024).

[25] Clarius Mobile Health Corp. *Clarius Ultrasound*. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K180799 (visited on 10/31/2024).

[26] Google. *Gemini*. URL: https://gemini.google.com/ (visited on 10/31/2024).

[27] *Google scholar*. URL: https://scholar.google.com/ (visited on 10/31/2024).

[28] Caption Health Inc. *Caption Interpretation Automated Ejection Fraction Software*. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN220063 (visited on 10/31/2024).

[29] Volpara Health Technologies Limited. *Volpara Imaging Software*. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K211279 (visited on 10/31/2024).

[30] Meta. *Llama*. URL: https://www.llama.com/ (visited on 10/31/2024).

[31] MITRE. *ATLAS*. URL: https://atlas.mitre.org/matrices/ATLAS (visited on 10/31/2024).

[32] OpenAI. *ChatGPT*. URL: https://chat.openai.com/ (visited on 10/31/2024).

[33] GE Medical Systems. *Auto Segmentation*. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm?ID=K230082 (visited on 10/31/2024).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 2 and 3 validate claims made in the Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section4 mentions the limitations of MedAIScout.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The details of the tool design in Section2 and the names of the devices used for evaluation (as mentioned in Section3) will enable a researcher to reproduce the experimental results. While the data used are publicly available, the code will be disclosed to interested researchers only.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: While the data used in this paper are publicly available, the code will be disclosed to interested researchers only.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: We do not train any model in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included a discussion on our evaluation results in Section3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details of compute resources have been mentioned in Section3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work presented in this paper conforms with all the NEURIPS Ethical Guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section1 highlights the impact of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: While the data used in this paper is publicly available, the code will only be shared with interested researchers after a thorough discussion on the plan of use.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make the best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data owners have been properly mentioned and cited at all appropriate places.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.