NFSU JOURNAL OF CYBER SECURITY & DIGITAL FORENSICS

NFSU – Journal of Cyber Security and Digital Forensics
Volume – 2, Issue – 2, December 2023
E – ISSN – 2583-7559

SAFEGUARDING AI-MEDIATED DIAGNOSES: A COMPREHENSIVE REVIEW OF CYBERSECURITY CHALLENGES AND SOLUTIONS IN LARGE LANGUAGE MODEL-ASSISTED MEDICAL APPLICATIONS

Param Ahir

Computer/IT Engineering
Gujarat Technological University
Ahmedabad, India
209999913009@gtu.edu.in

Mehul Parikh

Information Technology Department
L. D. College of Engineering
Ahmedabad, India
mehulparikh@ldce.ac.in

***Abstract*— This extensive review discusses the connection between large language models (LLMs) and medical imaging, exploring into the complex realm of cybersecurity challenges and solutions in the context of AI-assisted diagnoses. The paper provides a critical review of key studies that address the vulnerabilities and opportunities associated with the integration of LLMs in the analysis and interpretation of textual data related to medical images. Utilising a variety of research methods, such as qualitative analyses, quantitative studies, and ethical evaluations, this review combines findings from each study to provide a comprehensive overview of the present state of cybersecurity in LLM-assisted medical imaging. The comprehensive literature review establishes a basis for examining various obstacles, including potential risks to data privacy, malicious attacks, ensuring data integrity, and safeguarding network security. The paper explores various aspects related to cybersecurity measures, such as encryption protocols, strategies for defending against adversaries, and ethical considerations about the use of AI in healthcare. In this review, we seek to make a valuable contribution to the ongoing discussion surrounding the future of medical imaging at a time when large language models are becoming more prevalent.**

Keywords*—*large language model, medical imaging, artificial intelligence, cyber security

INTRODUCTION

In recent years, there has been a significant transformation in the field of healthcare due to the rapid rise of large language models (LLMs)[1]. The utilisation of cutting-edge models, driven by advancements in natural language processing and machine learning, has brought about a significant transformation in the analysis and interpretation of healthcare data. In recent years, the combination of LLMs and medical imaging technologies has opened up a world of opportunities. This exciting development allows for the thorough analysis of textual data associated with medical images, leading to improved diagnostic accuracy and more informed clinical decisions[2]. The emergence of large language models and federated learning[3] in healthcare can be attributed to the growing digitization of medical records and the rapid expansion of textual data. Models like GPT-4[4] and BERT[5], created for natural language understanding, have discovered fascinating uses in decoding intricate medical narratives, closing the divide between unstructured clinical text and quantitative data. The integration of LLMs with medical imaging technologies opens up new possibilities for comprehensive patient care by combining textual and visual information. With the ongoing advancements in medical imaging, the integration of LLMs allows for a comprehensive examination of both imaging data and the accompanying textual information[6].

In the following sections, we will thoroughly examine important studies, using various research methods such as qualitative analysis, quantitative assessment, and ethical evaluation. Our goal is to gain a deeper understanding of the cybersecurity issues and potential solutions in the ever-changing field of LLM-assisted medical imaging.

LLMS AND MEDICAL IMAGING: CURRENT ADVANCEMENTS

Significant advancements have been made in medical image analysis thanks to the emergence of large language models (LLMs), which have effectively bridged the gap between textual and visual data[7]. Utilising their advanced language processing abilities, LLMs enhance the understanding of medical images by placing them in the larger context of healthcare[8]. In particular, these models demonstrate exceptional proficiency in analysing radiological reports, clinical notes, and other textual data related to medical images. In this section, we will delve into the intricate world of LLM applications in medical image analysis. We will examine how these models bring a new level of clarity to the interpretation of complex imaging data, ultimately leading to a more comprehensive approach to diagnostic workflows. In this particular context, LLMs have a significant impact on automating the extraction of valuable information from medical texts. This enables a more thorough integration of clinical knowledge into the interpretation of imaging results.

- *Security Concerns and Challenges Related to LLM Integration*

The incorporation of LLMs with medical imaging technologies brings forth a range of cybersecurity concerns that require thoughtful examination. In this subsection, we will explore the security challenges related to LLM integration, highlighting the importance of addressing vulnerabilities to safeguard the confidentiality, integrity, and availability of patient data.

- Adversarial Attacks and Model Vulnerabilities[9]: LLMs, similar to any AI system, can be vulnerable to adversarial attacks. In this section, we will delve into situations where adversaries take advantage of weaknesses in LLMs to manipulate medical images or create deceptive textual results. These occurrences raise valid concerns about the reliability of these models in healthcare applications that require strong security measures.
- Privacy Concerns in Data[10]: The interaction between LLMs and sensitive medical data raises concerns about potential privacy risks. In this discussion, we will explore the importance of implementing strong privacy measures to protect patient information, particularly when it is being processed through LLMs. These measures are crucial to minimize the potential for unauthorised access or data breaches.

CRITICAL LITERATURE REVIEW: CYBERSECURITY IN LLM-ASSISTED MEDICAL IMAGING

Exploring the connection between large language models (LLMs) and medical imaging through a cybersecurity comprehensive literature review offers valuable insights into the current understanding, research methods, and wide array of discoveries. This section provides a comprehensive analysis of various studies that examine cybersecurity issues and potential solutions related to LLM-assisted medical imaging in the year 2023. Table 1 includes contribution, key findings, current limitation, and future scope.

Table 1 Survey of Existing Literature

Ref	**Contribution**	**Key Findings**	**Current Limitation**	**Future Scope**
[11]	Examines the applications of generative AI and LLMs in ophthalmology, focusing on their transformative potential in eye care.	Identifies challenges in integrating LLMs into clinical workflows and highlights the potential to enhance patient experiences.	Discusses constraints in ophthalmological examinations, privacy concerns, and the risk of false responses by LLMs.	Suggests further refinement in the role of LLMs in ophthalmology and emphasizes addressing ethical and legal concerns.
[12]	Provides an overview of safety and security threats and vulnerabilities of LLMs, focusing on criminal activities and AI alignment.	Highlights the misuse of LLMs for fraud, impersonation, malware generation, and the general problem of AI alignment.	Discusses the theoretical and practical limitations of LLM safety, including imperfect prevention measures.	Suggests focusing on potential future concerns related to LLM development and the evolution of LLM-enabled threats.
[13]	Evaluation of time, cost, and accuracy in developing LLM prompts for extracting clinical information from breast cancer patient reports.	High efficiency and accuracy (87.7% overall, 98.2% for lymphovascular invasion) in extracting information using LLM, time and cost-effectiveness compared to manual methods.	Clinical N stage does not strictly following AJCC staging, limitations in manual methods as control, and potential inaccuracies in extrapolating time and cost from a smaller sample.	Potential for broader application of LLM in medical data extraction and analysis need for validation with more data and improved accuracy measures.
[14]	Explores the use of LLMs (GPT-3.5 and GPT-4) for automating pragma-discursive corpus annotation, focusing on apology components in English.	GPT-4 showed superior performance compared to GPT-3.5 in annotating apologies, with accuracy approaching that of a human coder. Demonstrated the feasibility of using LLMs for efficient, scalable corpus annotation.	-	Suggests further exploration of LLMs in various linguistic annotation tasks, indicating the potential for broad application in corpus-based pragmatics and discourse analysis.
[15]	Introduces LLM-Assisted Content	LACA can significantly reduce the time	LACA may struggle with	Suggests further exploration of LLMs for

	Analysis (LACA) as a methodology to integrate LLMs into the deductive coding process, aligning with traditional content analysis while leveraging LLMs' capabilities.	required for deductive coding. GPT-3.5 achieved levels of agreement comparable to human coders in many cases, and the methodology helped identify areas where LLMs could or couldn't reliably code.	specific types of coding, particularly those involving formatting or character-level language aspects, due to limitations in LLMs' design.	deductive coding across various domains and datasets. Also, emphasizes the need for additional reporting and documentation standards for reproducibility and critique in research using LLMs.
[16]	Presents a framework that combines qualitative annotation and quantitative modelling, using LLMs to augment and automate tasks in the humanities and social sciences.	Demonstrates the applicability of LLMs across a wide range of tasks and disciplines. Shows how LLMs can augment traditionally qualitative areas, ensuring rigorous interpretations.	Discusses the need for rigorous statistical modelling to ensure the validity of inferred variables, highlighting a common oversight in quantitative research.	Emphasizes the potential for further exploration and application of LLMs in diverse research tasks within the humanities and social sciences, and the importance of integrating machine learning into research methodologies.

CYBERSECURITY CHALLENGES IN LLM-ASSISTED MEDICAL IMAGING

The incorporation of large language models (LLMs) with medical imaging technologies brings about numerous cybersecurity challenges that require careful attention to guarantee the secure and ethical implementation of these advanced AI systems. In this section, we delve into the significant cybersecurity challenges that arise in the context of LLM-assisted medical imaging. We conduct a comprehensive analysis of the potential risks to data privacy, the vulnerabilities of LLMs to adversarial attacks, the importance of maintaining data integrity, and the need for robust network security measures.

◦ *Data Privacy Concerns*

Robust encryption algorithms[17] are crucial for safeguarding the privacy of sensitive medical data during transmission and storage. State-of-the-art cryptographic methods like AES[18] and RSA[19] are utilised to secure medical images and their accompanying text data. By utilising these algorithms together, a robust communication channel is established, ensuring the utmost protection of patient data from any unauthorised access. Ensuring the confidentiality of medical image data during transmission from imaging devices to LLM processing centres is crucial in preventing interception by malicious entities. End-to-end encryption plays a vital role in maintaining the security of this data. The integration of AES encryption into a medical imaging platform has proven to be highly effective in mitigating the risk of unauthorised access during data transmission. This implementation has successfully safeguarded patient privacy and ensured compliance with regulat4.2 Exploring Adversarial Attacks on LLMs

◦ *Exploring Adversarial Attacks on LLMs*

It is crucial to employ advanced defence mechanisms due to the vulnerability of LLMs to adversarial attacks[20]. Exposing LLMs to manipulated data during training can significantly enhance their ability to handle adversarial inputs, making them more robust. Additionally, incorporating gradient masking techniques and employing adversarial detection algorithms[21] such as Feature Squeezing[22] enhances the model's resilience against adversarial perturbations. By integrating adversarial training into the model training pipeline, LLMs can effectively handle manipulated inputs, thereby minimizing the influence of adversarial attacks on the accuracy of medical diagnoses. A healthcare institution conducted a study on its LLM-based diagnostic system[23], implementing adversarial training. The results showed significant improvements in accuracy and the system's ability to withstand adversarial attempts to deceive it.

◦ *Maintaining Data Integrity*

Blockchain technology is utilised to ensure the integrity of medical imaging data processed by LLMs[24]. Every transaction or modification in the medical data is meticulously documented in a decentralised and tamper-evident blockchain. Smart contracts provide a mechanism to verify the integrity of the data, guaranteeing its immutability[25]. Introducing a cutting-edge system that utilizes blockchain technology ensures the utmost security and reliability when it comes to storing and accessing medical images. This innovative solution offers a clear and unchangeable log of every single interaction with the images, ensuring the integrity of patient data. A medical research institution has implemented a cutting-edge solution that utilizes blockchain technology to securely store and validate medical imaging data. This innovative approach guarantees the integrity and traceability of the data at every stage of its existence.

◦ *Enhancing Network Security*

Ensuring the establishment of secure communication channels is of utmost importance in safeguarding against any unauthorised access and interception of data. Transport

Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), are cryptographic protocols that ensure secure communication over a computer network[26]. By implementing TLS, the security of medical imaging data transmitted between devices and processing centres is enhanced. This encryption method effectively safeguards patient information, preventing unauthorised access and maintaining confidentiality. A healthcare network successfully implemented TLS for secure communication in their LLM-assisted medical imaging system, greatly enhancing the system's security and minimizing the potential for unauthorised access and data breaches.

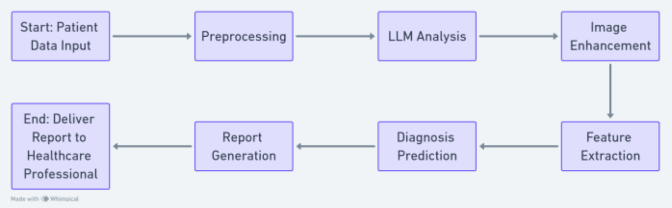

Figure.1 LLM Process flow for Medical Imaging

The following figure.1 depicts the extensive cybersecurity measures incorporated into the LLM-assisted medical imaging workflow. It highlights the algorithms and technologies utilised at each stage to address potential threats and vulnerabilities.

Enhancing Cybersecurity Measures in LLM-assisted Medical Imaging

When it comes to LLM-assisted medical imaging, it is crucial to prioritise cybersecurity measures to guarantee the ethical and secure implementation of these advanced technologies. In this section, we explore crucial aspects of cybersecurity measures, including encryption and data transmission, robustness and adversarial defence, explainability and transparency, and ethical considerations.

◦ *Encryption and Data Transmission*

Ensuring the utmost security of medical imaging data transmission and storage is of paramount importance, and one way to achieve this is through the implementation of strong encryption protocols like AES and RSA. Ensuring the privacy of patient information is a crucial aspect of data transmission. Ensuring the security of communication channels is crucial in safeguarding medical imaging data from unauthorised access and interception. This can be achieved by implementing robust protocols like TLS and SSL.

◦ *Ensuring Resilience and Protection Against Adversaries*

Incorporating adversarial training into the model development process strengthens the resilience of LLMs against adversarial attacks. Exposure to manipulated data during training enhances the models' ability to handle unexpected inputs. By incorporating sophisticated defence techniques such as gradient masking and feature squeezing, LLMs can effectively safeguard against adversarial perturbations, ensuring the precision and dependability of medical diagnoses.

◦ *Increasing Awareness*

Emphasizing the importance of interpretable LLM models enhances our comprehension of their decision-making processes. The transparency of this process not only fosters trust in the clinical realm but also helps to uncover any possible security weaknesses. By integrating explanatory interfaces, healthcare professionals can gain a deeper understanding of LLM outputs. This allows them to validate the decisions made by the model, creating a collaborative and secure diagnostic environment.

◦ *Ethical Considerations*

Ensuring ethical standards are met requires obtaining informed consent from patients for the utilisation of medical data in LLM-assisted imaging. Effective communication about data usage and its potential consequences is crucial for establishing patient trust. It is crucial to maintain fairness and minimize biases in LLMs. Ensuring fair healthcare outcomes for all patients requires a continuous focus on minimizing biases in algorithms and upholding ethical standards[27]. The implementation of these cybersecurity measures establishes a strong and morally upright basis for incorporating LLMs into medical imaging. By emphasizing the importance of encryption, ensuring strong and reliable systems, promoting transparency, and adhering to ethical standards, the healthcare community can effectively address the challenges presented by these cutting-edge technologies while safeguarding patient privacy and well-being.

Future Scope

The field of large language model (LLM)-assisted medical imaging is constantly changing, offering possibilities for progress in cybersecurity, collaborative initiatives, and the ethical development of AI.

- **Advances in LLM-based Cybersecurity:** Future advancements could utilise federated learning techniques to improve LLM-based cybersecurity. Through the collaborative training of models on decentralised devices, this approach enhances security and maintains the confidentiality of data. The progress in defending against adversaries will be crucial. Utilising techniques like ensemble methods and reinforcement learning integration can enhance the resilience of LLMs in medical imaging, safeguarding against adversarial attacks and ensuring their robustness.
- **Collaborative Efforts:** Encouraging partnerships between academic institutions and industry will be crucial. Collaborative research endeavours have the potential to foster creativity and ingenuity by merging the knowledge of scholars with practical problems, aiming to tackle the ever-evolving cybersecurity concerns in LLM-assisted medical imaging. Collaborating with experts from healthcare, technology, and regulatory fields will enable a comprehensive approach to cybersecurity. Collaborative frameworks that involve healthcare providers, technology developers, and policymakers can establish unified standards and practices.
- **The Ethical Development of AI:** It is anticipated that in the future, there will be a push towards implementing standardised practices for explainable AI. By establishing

precise criteria for model interpretability, transparency is achieved in LLM-based medical imaging. This approach effectively tackles ethical concerns and fosters a greater sense of trust. It is crucial to prioritise the establishment of community-driven ethical guidelines for AI in healthcare. Engaging healthcare professionals, ethicists, and patient advocates in the development of ethical frameworks guarantees a well-rounded and inclusive approach to AI advancement.

Embracing these future directions and innovations will be crucial in shaping a secure, collaborative, and ethically grounded landscape for LLM-assisted medical imaging as the field progresses. By remaining at the cutting edge of cybersecurity advancements, promoting collaboration, and adhering to ethical principles, the healthcare community can fully utilise LLMs while guaranteeing their responsible and secure implementation.

CONCLUSION

In conclusion, the combination of extensive language models (LLMs) and medical imaging presents a vast range of possibilities, necessitating a thorough examination of cybersecurity obstacles. Understanding the various aspects of this field, such as encryption, adversarial defence, transparency, and ethical guidelines, is crucial for ensuring the ethical and secure deployment of LLMs. In the future, the progress in LLM-based cybersecurity, collective initiatives, and the establishment of ethical AI standards will have significant impacts. Ensuring a harmonious blend of innovation and ethical integrity is crucial for LLMs in medical imaging to make significant contributions to advanced diagnostics, all while placing utmost importance on patient privacy and data security in the healthcare sector.

REFERENCES

[1] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940.

[2] Li, Y., Wang, H., & Luo, Y. (2020, December). A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1999-2004). IEEE.

[3] Ahir, P., & Parikh, M. Cyber Security Concerns and Mitigation Strategies in Federated Learning: A Comprehensive Review.

[4] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.

[5] Müller, M., Salathé, M., & Kummervold, P. E. (2023). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. Frontiers in Artificial Intelligence, 6, 1023281.

[6] Ahira, P., & Diwanjia, H. M. Analysis of Visual Question Answering Algorithms with attention model.

[7] Ahir, P., & Parikh, M. (2023, January). A Review of Recent Advancements in Infant Brain MRI Segmentation Using Deep Learning Approaches. In International Conference on Smart Trends in Computing and Communications (pp. 439-452). Singapore: Springer Nature Singapore.

[8] Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine, 6(1), 120.

[9] Kumar, A., Agarwal, C., Srinivas, S., Feizi, S., & Lakkaraju, H. (2023). Certifying llm safety against adversarial prompting. arXiv preprint arXiv:2309.02705.

[10] Montagna, S., Ferretti, S., Klopfenstein, L. C., Florio, A., & Pengo, M. F. (2023, September). Data decentralisation of llm-based chatbot systems in chronic disease self-management. In Proceedings of the 2023 ACM Conference on Information Technology for Social Good (pp. 205-212).

[11] Betzler, B. K., Chen, H., Cheng, C. Y., Lee, C. S., Ning, G., Song, S. J., ... & Wong, T. Y. (2023). Large language models and their impact in ophthalmology. The Lancet Digital Health, 5(12), e917-e924.

[12] Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. arXiv preprint arXiv:2308.12833.

[13] Choi, H. S., Song, J. Y., Shin, K. H., Chang, J. H., & Jang, B. S. (2023). Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiation Oncology Journal, 41(3), 209.

[14] Yu, D., Li, L., Su, H., & Fuoli, M. (2023). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. International Journal of Corpus Linguistics.

[15] Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. arXiv preprint arXiv:2306.14924.

[16] Karjus, A. (2023). Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. arXiv preprint arXiv:2309.14379.

[17] Adeniyi, A. E., Abiodun, K. M., Awotunde, J. B., Olagunju, M., Ojo, O. S., & Edet, N. P. (2023). Implementation of a block cipher algorithm for medical information security on cloud environment: using modified advanced encryption standard approach. Multimedia Tools and Applications, 1-15.

[18] Adityaa, G., & Lavanya, V. (2023, January). A Decentralized Storage System for 3D Medical Data with Dynamic AES and AES-GCM Encryption. In Recent Developments in Electronics and Communication Systems: Proceedings of the First International Conference on Recent Developments in Electronics and Communication Systems (RDECS-2022) (Vol. 32, p. 269). IOS Press.

[19] Shivaramakrishna, D., & Nagaratna, M. (2023). A novel hybrid cryptographic framework for secure data storage in cloud computing: Integrating AES-OTP and RSA with adaptive key management and Time-Limited access control. Alexandria Engineering Journal, 84, 275-284.

[20] Schwinn, L., Dobre, D., Günnemann, S., & Gidel, G. (2023). Adversarial attacks and defenses in large language models: Old and new threats. arXiv preprint arXiv:2310.19737.

[21] Meenakshi, K., & Maragatham, G. (2023). An Optimised Defensive Technique to Recognize Adversarial Iris Images Using Curvelet Transform. Intelligent Automation & Soft Computing, 35(1).

[22] Cha, J., Kang, W., Mun, J., & Roh, B. (2023). Honeybee: Locality-enhanced Projector for Multimodal LLM. arXiv preprint arXiv:2312.06742.

[23] Hur, K., Oh, J., Kim, J., Kim, J., Lee, M. J., Cho, E., ... & Choi, E. (2023). GenHPF: General Healthcare Predictive Framework for Multi-task Multi-source Learning. IEEE Journal of Biomedical and Health Informatics.

[24] Cabrera, J., Loyola, M. S., Magaña, I., & Rojas, R. (2023, June). Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In International

Work-Conference on Bioinformatics and Biomedical Engineering (pp. 313-326). Cham: Springer Nature Switzerland.