Balancing Privacy and Explainability in Healthcare AI through Secure Multi-Party Computation and Local Interpretable Model-Agnostic Explanations

Michael Chen
School of Health Professions, University of Texas Medical Branch
m.chen@utmb.edu

Abstract

The rapid integration of artificial intelligence within healthcare systems has precipitated a fundamental tension between the necessity for stringent data privacy and the clinical requirement for model interpretability. While advanced deep learning architectures offer unprecedented diagnostic accuracy, their inherent opacity often clashes with the ethical and legal mandates for transparency in medical decision-making. Simultaneously, the sensitive nature of patient health information necessitates robust protection against data leakage, often complicating the deployment of centralized transparency tools. This research paper explores a system-level synthesis of Secure Multi-Party Computation and Local Interpretable Model-Agnostic Explanations to bridge this gap. By leveraging cryptographic protocols to facilitate collaborative computation without exposing underlying datasets and integrating post-hoc explanation frameworks to clarify model behavior, this study proposes a robust architectural paradigm for trustworthy healthcare AI. The discussion emphasizes the structural trade-offs between computational overhead and clinical utility, the governance frameworks required for decentralized auditing, and the policy implications for data sovereignty in international medical research. Ultimately, the paper argues that the sustainability of healthcare AI depends on a socio-technical approach that treats privacy and explainability as mutually reinforcing objectives rather than zero-sum constraints, providing a comprehensive roadmap for deploying secure and interpretable diagnostic infrastructures in diverse clinical environments.

Keywords:
Healthcare Artificial Intelligence, Secure Multi-Party Computation, Local Interpretable Model-Agnostic Explanations, Data Privacy, Model Interpretability, Socio-Technical Systems

1. Introduction

The modern healthcare landscape is undergoing a profound digital transformation characterized by the convergence of high-performance computing, large-scale data aggregation, and sophisticated machine learning models. As medical institutions transition from traditional diagnostic workflows to AI-assisted decision support systems, the promise of

precision medicine becomes increasingly tangible. However, this evolution is shadowed by two critical challenges that threaten the widespread adoption and ethical integrity of these technologies: the preservation of patient privacy and the demand for algorithmic interpretability. The clinical environment is unique in its requirements for trust; a physician cannot ethically or legally act on a recommendation from a "black box" model without a fundamental understanding of the underlying rationale. Yet, the data required to train these models—electronic health records, genomic sequences, and high-resolution imaging—is subject to some of the most rigorous privacy regulations globally, such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe.

The tension between these two requirements often creates a technical stalemate. On one hand, enhancing explainability frequently involves analyzing raw data or revealing internal model weights, both of which can inadvertently expose sensitive information or proprietary intellectual property. On the other hand, traditional privacy-preserving techniques, such as data anonymization or differential privacy, can degrade the quality of the data to the point where explanation tools become unreliable or inaccurate. This paper addresses this systemic conflict by investigating the integration of Secure Multi-Party Computation (SMPC) and Local Interpretable Model-Agnostic Explanations (LIME). SMPC offers a framework where multiple parties can jointly compute a function over their inputs while keeping those inputs private, effectively allowing for the training and querying of models across institutional silos without centralized data storage [14]. When paired with LIME, which provides interpretable insights into individual predictions by perturbing input data and observing changes in output, a dual-layer defense and transparency mechanism emerges.

This research adopts a systems-oriented perspective, analyzing how these technologies interact within the broader socio-technical infrastructure of contemporary medicine. Rather than focusing solely on the technical nuances of cryptographic primitives or perturbation algorithms, the discussion examines the governance, deployment, and sustainability of such integrated systems. It evaluates the trade-offs in computational efficiency, the impact on clinical latency, and the shifting responsibilities of healthcare providers in a decentralized AI ecosystem. By synthesizing these perspectives, the paper aims to provide a high-level theoretical and practical framework for the next generation of healthcare AI, ensuring that technological progress does not come at the expense of patient rights or clinical accountability [32].

2. The Privacy-Explainability Paradox in Clinical AI

The pursuit of high-fidelity AI in healthcare has historically relied on the centralization of massive datasets. Centralized models benefit from a global view of patient populations, enabling the detection of subtle patterns that might be missed in localized studies. However, the centralization of medical data represents a significant security risk and a policy hurdle. Data breaches in the medical sector are uniquely damaging because health data is immutable and deeply personal. Consequently, many institutions are hesitant to share data, leading to

"data silos" that limit the generalizability of AI models. Privacy-preserving technologies have been proposed to mitigate these risks, but they often introduce a secondary problem: the degradation of model transparency. If a model is trained on encrypted or noise-injected data, the path to understanding its decisions becomes significantly more complex. This creates the privacy-explainability paradox, where the more one protects the data, the harder it becomes to explain why a particular diagnostic conclusion was reached.

From a structural standpoint, this paradox is rooted in the information loss associated with many privacy techniques. Differential privacy, for instance, adds statistical noise to the data to prevent the identification of specific individuals. While this is effective for privacy, it can obscure the very features that an explanation tool like LIME needs to identify as significant. If the explanation tool identifies a "noisy" feature as the primary driver of a prediction, the resulting clinical insight is at best useless and at worst dangerous. Conversely, explanation methods that require deep access to the model's architecture or the training set to provide global interpretations can act as a side-channel for data extraction. Sophisticated attackers can sometimes reconstruct training samples by observing the explanations generated by the system, a phenomenon known as an inversion attack [22].

To resolve this, healthcare systems must move toward architectures that decoupling data access from model interrogation. The goal is to create a system where the "truth" of the explanation is preserved without the "visibility" of the underlying sensitive data. This requires a shift from viewing privacy and explainability as separate modular additions to viewing them as an integrated feature of the system's core architecture. This integration involves rethinking the entire data lifecycle, from the point of collection at the bedside to the generation of a post-hoc explanation in the clinician's office. The complexity of medical decision-making, which involves multifaceted variables and high-stakes outcomes, necessitates that this integrated approach be both robust and flexible enough to handle various types of clinical data, from tabular laboratory results to complex spatio-temporal imaging [9].

3. Secure Multi-Party Computation: Architecting Decentralized Trust

Secure Multi-Party Computation represents a paradigm shift in how healthcare data is handled. In a traditional centralized model, trust is placed in a single entity—the data center or the cloud provider. In an SMPC framework, trust is distributed across multiple participants. The fundamental principle is that a computation is divided into secret shares, which are distributed among different servers or institutions. No single participant can see the original data, yet collectively, they can perform complex mathematical operations to train a model or generate an inference. In the context of healthcare, this allows a network of hospitals to collaborate on a diagnostic model for a rare disease without any patient records ever leaving the local firewalls of their respective institutions.

The implementation of SMPC in a large-scale healthcare system involves significant structural trade-offs. The most prominent of these is the communication and computational overhead. Cryptographic protocols require multiple rounds of interaction between

participating nodes, which can lead to latencies that are incompatible with real-time clinical needs, such as emergency room triage. However, for elective diagnostics, oncology planning, or long-term risk stratification, the latency of SMPC is a manageable trade-off for the unprecedented level of data security it provides. Furthermore, the sustainability of an SMPC-based infrastructure depends on the stability and bandwidth of the underlying network. As medical institutions are often disparate in their technological maturity, the system must be designed to be resilient to node failures or slow connections, requiring sophisticated load-balancing and fault-tolerant protocols [18].

Governance and policy also play a critical role in the deployment of SMPC. Unlike centralized databases where a single legal agreement might suffice, SMPC requires a multilateral governance framework. Institutions must agree on the protocols, the specific functions to be computed, and the methods for auditing the system without compromising the privacy of the secret shares. This necessitates a new form of digital medical ethics where the focus shifts from "who owns the data" to "who is authorized to participate in the computation." Such a shift aligns with the growing movement toward data sovereignty, where patients and local institutions retain control over their digital assets while still contributing to the global pool of medical knowledge. The use of SMPC thus provides a technical foundation for a more democratic and secure healthcare research ecosystem [1].

4. Local Interpretable Model-Agnostic Explanations: Bridging the Human-AI Gap

While SMPC secures the data, Local Interpretable Model-Agnostic Explanations (LIME) address the human element of the AI system. LIME functions by taking an individual instance—for example, a specific patient's diagnostic report—and perturbing the features of that instance to see how the model's prediction changes. By observing these fluctuations, LIME builds a simpler, interpretable model (like a linear regression) that approximates the complex "black box" model locally around that specific patient. This provides the clinician with a set of "reasons" for the prediction, such as "elevated glucose levels" or "specific patterns in the chest X-ray," which can then be validated against clinical expertise.

The model-agnostic nature of LIME is particularly valuable in healthcare, where the state-of-the-art model for a specific task might change frequently. Whether the underlying architecture is a deep neural network, a random forest, or a gradient-boosted tree, LIME can provide consistent, human-readable explanations. This consistency is vital for clinical training and for maintaining a standardized level of care across different technological implementations. However, the use of LIME in a clinical setting is not without its challenges. The quality of the explanation is highly dependent on the "neighborhood" of perturbations. If the perturbations are too small, they may not capture the model's behavior; if they are too large, they may veer into regions of the data space that are clinically impossible, leading to nonsensical explanations.

Integrating LIME into a secure framework requires ensuring that the perturbation process itself does not become a vector for privacy leakage. In an SMPC environment, the generation

of explanations must also be performed securely. This means that the perturbed samples and their corresponding model outputs must remain within the encrypted domain until the final, simplified explanation is generated. This "Secure LIME" approach ensures that even the process of explaining a decision respects the privacy constraints of the original data. From a socio-technical perspective, this reinforces the clinician's role as the ultimate arbiter of the AI's output. By providing a clear rationale, LIME empowers doctors to reject AI suggestions that are based on spurious correlations, thereby enhancing the overall robustness and safety of the healthcare system [27].

5. System Synthesis: Integrating Privacy and Interpretability

The core contribution of this research is the conceptual and structural synthesis of SMPC and LIME into a unified healthcare AI architecture. This integrated system operates through a multi-layered process. At the foundation, the data layer remains decentralized, with patient records residing in local hospital databases. The computation layer utilizes SMPC protocols to perform model training and inference. When a clinician requests a prediction for a new patient, the system generates the result through secure protocols. Simultaneously, the explanation layer activates, using the same secure framework to generate a local explanation via LIME. The final output delivered to the clinician is a dual-component package: the diagnostic prediction and the interpretable rationale behind it.

This synthesis addresses the structural trade-offs of both technologies. SMPC provides the "where" and "how" of secure processing, while LIME provides the "why." Together, they create a system that is greater than the sum of its parts. For instance, the explainability provided by LIME can serve as an internal auditing tool for the SMPC process. If the explanations consistently point to irrelevant or biased features, it may indicate a problem with the training data or a flaw in the cryptographic implementation that is distorting the model's learning. This feedback loop is essential for the long-term sustainability and reliability of AI in high-stakes environments like oncology or cardiology.

From a deployment perspective, such a system requires a robust socio-technical infrastructure. This includes not only the hardware and software but also the institutional policies, training programs for medical staff, and legal frameworks for liability. If an AI system provides a secure, explained prediction that turns out to be incorrect, the traditional lines of malpractice and product liability become blurred. The infrastructure must, therefore, include comprehensive logging and auditing capabilities that can reconstruct the decision-making process for legal review without compromising patient privacy. This level of systemic transparency is crucial for building the public trust necessary for the widespread adoption of AI in public health [30].

6. Deployment Challenges and Computational Feasibility

Deploying an integrated SMPC-LIME system in a real-world healthcare environment involves overcoming significant technical and logistical hurdles. The first is the sheer scale of

modern medical data. Genomic datasets, for example, can be several terabytes in size. Running SMPC on such massive volumes of data requires specialized hardware, such as Trusted Execution Environments or high-performance cryptographic accelerators. Furthermore, the iterative nature of LIME—which requires hundreds or thousands of model queries per explanation—multiplies the computational cost when those queries must be performed through an SMPC protocol. This necessitates the development of more efficient cryptographic primitives and optimized explanation algorithms specifically designed for the secure domain.

Another major challenge is the heterogeneity of healthcare data. Medical records are a mix of structured data, unstructured text, and multi-dimensional images. Standardizing these formats across different institutions so they can be processed by a unified SMPC-LIME system is a massive undertaking in data engineering. It requires the adoption of common data models and interoperability standards, such as Fast Healthcare Interoperability Resources. Without this standardization, the "secret shares" used in SMPC will not align, and the explanations generated by LIME will be inconsistent across different hospital systems. The robustness of the system is therefore tied directly to the quality of the underlying data governance and standardization efforts.

Sustainability also concerns the energy and resource consumption of these systems. Cryptographic computations are inherently more resource-intensive than clear-text processing. As healthcare systems strive to reduce their carbon footprint, the environmental impact of large-scale SMPC clusters must be considered. This points toward a need for "green" cryptographic solutions and edge-computing architectures that can perform some of the secure computation closer to the data source, reducing the energy cost of data transmission. The feasibility of these systems is not just a matter of "can we build it" but "can we afford to maintain it" over the long term in a resource-constrained healthcare environment [5].

7. Fairness, Bias, and Algorithmic Governance

A critical aspect of healthcare AI is the mitigation of bias and the promotion of fairness. AI models are prone to inheriting the biases present in their training data, which can lead to disparities in care for marginalized populations. In a centralized system, detecting this bias is difficult; in a decentralized, privacy-preserving system like the one proposed here, it is even more challenging. Because no single entity has access to the full dataset, it is hard to perform the standard statistical checks for demographic parity or equalized odds. However, the integration of LIME offers a potential solution. By analyzing the explanations generated for different demographic groups, researchers can identify if the model is using inappropriate proxies (like zip code or insurance type) for its clinical predictions.

Governance of such a system must be multi-scalar, involving local clinical oversight, institutional review boards, and national regulatory bodies. A decentralized auditing framework can be established where "auditor" nodes participate in the SMPC process to verify fairness metrics without ever seeing individual patient data. This allows for a

continuous monitoring of the system's performance and ethical alignment. Furthermore, the transparency provided by LIME is a prerequisite for informed consent. Patients should have the right to know not only that an AI was used in their care, but also the general basis on which the AI made its decision. This aligns with the "right to an explanation" enshrined in modern data protection laws.

The policy implications extend to the international level. Collaborative medical research often crosses national borders, bringing disparate legal regimes into conflict. A secure, explainable AI infrastructure can act as a "neutral ground" for international collaboration. By providing technical guarantees of privacy and transparency, it can facilitate the sharing of medical insights between countries that might otherwise be prevented from sharing data due to political or legal constraints. This has profound implications for global health security, particularly in the rapid response to pandemics or the study of rare genetic conditions. The system thus becomes a tool for both technological innovation and diplomatic cooperation [19].

8. Future Directions and Forward-Looking Perspectives

Looking toward the future, the integration of SMPC and LIME is likely to be augmented by emerging technologies like federated learning and blockchain. Federated learning could provide a more efficient way to train models across decentralized nodes, while SMPC could be reserved for the final, sensitive stages of model aggregation or for the explanation process itself. Blockchain or distributed ledger technology could provide an immutable audit trail of the SMPC-LIME operations, ensuring that the history of every clinical prediction and its explanation is securely recorded for future review. This convergence of technologies points toward a "trustless" medical infrastructure where security and transparency are guaranteed by mathematics and code rather than institutional promises.

Furthermore, the evolution of human-computer interaction in medicine will necessitate more sophisticated explanation interfaces. Instead of static text or heatmaps, future LIME-based systems might offer interactive, conversational explanations, allowing clinicians to ask follow-up questions like "What if the patient's blood pressure was lower?" in a secure environment. This would turn the AI from a simple diagnostic tool into a true collaborative partner. Research into "Explainable AI for Healthcare" must also consider the cognitive load on the clinician. Explanations should be concise and prioritized based on clinical relevance to avoid "alert fatigue" and ensure that the most critical insights are acted upon promptly.

Finally, the long-term robustness of these systems will depend on their ability to adapt to changing medical knowledge. A model trained today may be obsolete in two years as new treatments and diagnostic criteria emerge. The SMPC-LIME architecture must be designed to be modular and upgradeable, allowing for the periodic retraining of models and the refinement of explanation parameters without disrupting the ongoing clinical operations. This requires a rethink of the medical software lifecycle, moving away from "set it and forget it" deployments to a model of continuous evaluation and evolution. The future of healthcare AI

lies in systems that are not only secure and explainable but also inherently resilient and adaptive [2].

9. Conclusion

The integration of Secure Multi-Party Computation and Local Interpretable Model-Agnostic Explanations represents a vital step forward in the development of trustworthy healthcare AI. By providing a technical solution to the privacy-explainability paradox, this architecture allows for the collaborative power of large-scale data analysis without sacrificing individual patient rights or clinical transparency. The structural discussion provided in this paper highlights that while the computational and logistical challenges are significant, they are not insurmountable. The trade-offs in latency and resource consumption are outweighed by the benefits of a secure, fair, and interpretable diagnostic infrastructure.

Ultimately, the success of such systems depends on a holistic, socio-technical approach. Technical innovation must be matched by robust governance, standardized data practices, and a commitment to medical ethics. As AI becomes more deeply embedded in the fabric of clinical care, the ability to explain "why" a decision was made, and the guarantee that the data used was handled with the utmost security, will be the cornerstones of patient trust. This paper provides the conceptual roadmap for achieving that vision, ensuring that the AI-driven future of medicine is as transparent and secure as it is effective. The path forward requires ongoing interdisciplinary collaboration between computer scientists, clinicians, ethicists, and policymakers to refine these systems for the diverse and complex realities of global healthcare.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

2. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

3. Ahmad, M. A., Eckert, C., & Tancik, M. (2018). Interpretable machine learning in healthcare. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 559-560.

4. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 1-9.

5. Beaulieu-Jones, B. K., Yuan, W., Brat, G. A., Beam, A. L., Weber, G., Wyatt, M., ... & Kohane, I. S. (2019). Machine learning for patient risk stratification: Standing on, or

looking over, the shoulders of clinicians? npj Digital Medicine, 2(1), 1-6.

6. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1.

7. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191.

8. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730.

9. Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. International Conference on Machine Learning, 883-892.

10. Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: Graph-based attention model for healthcare representation learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 787-795.

11. Dwork, C. (2008). Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 1-19.

12. Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taneja, A. (2019). Explainable AI in industry. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3203-3204.

13. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings, 2020, 191.

14. Goldreich, O. (2004). Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press.

15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), 1-42.

16. He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical

implementation of artificial intelligence in echocardiography. Nature Reviews Cardiology, 16(5), 290-297.

17. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 1-9.

18. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

19. Kundu, S. (2019). AI in medicine: The 4th industrial revolution. International Journal of Surgery, 68, 82-84.

20. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

21. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Artificial Intelligence and Statistics, 1273-1282.

22. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. 2017 IEEE Symposium on Security and Privacy (SP), 19-38.

23. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

24. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

25. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 234-241.

26. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer Nature.

27. Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham Law Review, 87, 1085.

28. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

29. Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199-2200.

30. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

31. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689.

32. Wang, F., Casalino, L. P., & Khullar, D. (2019). Deep learning in medicine—promise, progress, and dangers. JAMA Internal Medicine, 179(3), 293-294.

33. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

34. Zyskind, G., & Nathan, O. (2015). Decentralizing privacy: Using blockchain to protect personal data. 2015 IEEE Security and Privacy Workshops, 180-184.