

Energy-Efficient Privacy-Preserving Neural Networks for Resource-Constrained IoT Devices Using Secure Computation

Gioacchino Cilibrasi

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
gioacchino.cilibrasi387@colostate.edu

Iwo Kniec

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
iwo1994@uab.edu

Przemyslaw Kowala

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
przemyslaw21@ucf.edu

Abstract

Artificial intelligence at the edge is increasingly embedded in consumer, industrial, medical, and civic infrastructures, yet many internet-of-things devices operate under severe limits in energy, memory, bandwidth, and thermal headroom. Running neural networks locally can reduce latency and communication, but it also exposes sensitive data to device compromise and raises questions about whether privacy can be preserved without sacrificing operational sustainability. This paper examines energy-efficient privacy-preserving neural networks for resource-constrained internet-of-things devices through the lens of secure computation. Rather than treating cryptography, machine learning optimization, and infrastructure governance as separate domains, the analysis frames them as coupled design choices that shape system architecture, cost, robustness, and fairness. The discussion considers secure multi-party computation, homomorphic encryption, secret sharing, differential privacy, and federated learning, with attention to their energy profiles and deployment feasibility in edge and gateway environments. It then analyzes structural trade-offs among privacy guarantees, inference accuracy, latency, battery life, and communication overhead. The paper argues that practical deployments require hybrid protocols, model compression, reduced activation complexity, secure offloading, and policy mechanisms that clarify accountability across device manufacturers, network operators, cloud providers, and public institutions. Case illustrations from smart health, industrial monitoring, and urban sensing show that privacy and energy efficiency are not merely technical constraints but also governance and sustainability challenges. The conclusion outlines a research agenda for auditable, inclusive, and resource-aware secure neural computation at the edge.

Keywords

secure computation; privacy-preserving machine learning; internet of things; energy efficiency; edge intelligence; federated learning; homomorphic encryption; differential privacy; governance; sustainability.

1. Introduction

The rapid expansion of internet-of-things systems has moved neural network inference from centralized data centers into homes, factories, hospitals, vehicles, and public spaces. This shift promises lower latency, reduced backhaul traffic, and improved autonomy when connectivity is intermittent. However, it also creates a difficult sociotechnical problem: the data processed by these devices are often personal, proprietary, or safety-critical, while the devices themselves are constrained by small batteries, limited memory, weak processors, and unreliable network links. Secure computation offers a principled response by allowing computation to occur without exposing raw inputs to untrusted parties, but its cryptographic overhead can be incompatible with the energy budgets of constrained devices. The central challenge is therefore not simply whether privacy can be achieved, but whether it can be achieved in a manner that is energy-efficient, robust, fair, and deployable at scale.

Early work on secure computation established the theoretical basis for evaluating functions over private inputs [1], and subsequent research extended these ideas to general multi-party protocols and composable security models [2], [3], [4]. In parallel, privacy-preserving machine learning has developed specialized architectures for neural network inference using homomorphic encryption, secret sharing, and hybrid protocols [5], [6], [7], [8], [9], [10]. These advances demonstrate that private inference is technically possible, yet most systems still assume relatively powerful servers or gateways. Resource-constrained internet-of-things devices introduce an additional layer of complexity because energy consumption is not only an operational cost but also a determinant of device lifetime, maintenance schedules, and environmental impact. A privacy mechanism that drains batteries or forces frequent battery replacement may be socially and economically unacceptable, even if it is cryptographically sound.

This paper argues that energy-efficient privacy-preserving neural networks for constrained internet-of-things devices must be understood as an infrastructure design problem. The analysis examines how secure computation protocols, neural network architectures, deployment topologies, and governance arrangements interact. It emphasizes structural trade-offs rather than isolated algorithmic improvements, because choices about where computation occurs, which party holds which secret, and how models are compressed or partitioned determine both privacy and energy outcomes. The discussion also considers robustness against adversarial manipulation, fairness across populations with different device capabilities, and sustainability in relation to e-waste and energy use. The remainder of the paper reviews foundational work, describes system architecture and threat models, analyzes energy constraints, evaluates privacy-preserving strategies, examines optimization trade-offs, discusses governance and policy, and concludes with a forward-looking research agenda.

2. Background and Related Work

Secure computation began with foundational protocols for private two-party and multi-party evaluation [1], [2]. These works showed that arbitrary functions can be computed without revealing inputs beyond what the output implies, but they also revealed substantial communication and computation costs. Later research introduced circuit randomization, universally composable security, and efficient multiparty computation from somewhat homomorphic encryption, expanding the practical scope of secure computation [3], [4], [5]. These contributions remain important for internet-of-things systems because they clarify the trust assumptions, adversarial models, and composability guarantees that must be considered when multiple devices, gateways, and cloud services participate in a single workflow.

Homomorphic encryption provides another route to privacy-preserving inference by allowing computation directly on encrypted data. CryptoNets demonstrated that neural networks could be evaluated over encrypted inputs with meaningful throughput and accuracy [6]. Subsequent work on oblivious neural network predictions, hybrid secure computation, and low-latency frameworks has refined the balance between cryptographic security and practical performance [7], [8], [9], [10]. These systems often rely on secure multi-party computation for non-linear operations, while using homomorphic encryption or secret sharing for linear layers. The resulting architectures are powerful but sensitive to network latency, model size, and activation functions, all of which become critical in resource-constrained environments.

Neural network efficiency research has produced complementary techniques such as pruning, quantization, Huffman coding, and compact convolutional architectures [11], [12]. These methods reduce memory footprint, computation, and energy use, making local inference more feasible. However, compression can interact with secure computation in non-obvious ways. For example, quantization may reduce communication in secret-sharing protocols but can complicate numerical precision, while pruning may remove parameters that were important for maintaining accuracy under encrypted or masked computation. Similarly, federated learning and secure aggregation reduce the need to centralize raw data, but they introduce communication rounds and coordination costs that can be burdensome for battery-powered devices [13], [14]. Differential privacy offers formal privacy guarantees and can be combined with federated learning, yet it introduces a privacy-utility trade-off that must be tuned carefully in edge settings [15], [18].

3. System Architecture and Threat Model

A practical architecture for energy-efficient private neural computation in internet-of-things environments typically involves multiple tiers: sensing devices, local edge gateways, and remote cloud or institutional servers. Sensing devices may perform lightweight feature extraction or run compact models, while gateways aggregate, pre-process, or execute secure computation protocols on behalf of devices with limited power. Remote servers may provide model updates, certificate management, or heavy cryptographic operations. This tiered structure is not merely a performance convenience; it defines the trust boundaries that secure computation must protect. If a gateway is compromised, it may learn intermediate representations, timing patterns, or metadata unless the protocol is designed to limit leakage [4], [9].

Threat models for such systems must account for honest-but-curious adversaries, malicious participants, colluding servers, compromised devices, and side-channel observers. Honest-but-curious adversaries follow the protocol but attempt to infer private information from messages, while malicious adversaries may deviate arbitrarily. In internet-of-things deployments, physical access to devices makes hardware tampering and side-channel extraction realistic concerns. Secure computation can protect against some of these threats, but only if the protocol assumptions match the deployment environment. For example, a two-party protocol may assume that one server is trusted not to collude with another, but in a smart-city deployment, the same municipal operator may control both parties. Governance therefore becomes inseparable from architecture, because trust assumptions must be documented, audited, and aligned with institutional accountability [1], [2], [4].

The placement of secure computation also affects energy use. Running cryptographic protocols on every constrained device may be infeasible, while offloading all private computation to a gateway may create a central point of trust. Hybrid designs can distribute

work according to device capability and privacy sensitivity. For instance, a wearable health monitor might encrypt or secret-share only the most sensitive features, while a gateway performs the bulk of private inference and returns a result. This approach can reduce energy consumption on the device, but it increases dependence on the gateway and network availability. The architecture must therefore specify failure modes, fallback behavior, and transparency requirements, especially when clinical, financial, or safety-critical decisions are involved [7], [8], [10].

4. Energy and Resource Constraints in IoT

Energy constraints in internet-of-things devices arise from limited battery capacity, thermal dissipation limits, and the cost of wireless communication. Computation, memory access, and radio transmission all consume energy, but communication often dominates in distributed machine learning workflows. Sending raw data to a cloud server can be energy-intensive, yet local private inference can also be expensive if cryptographic operations require repeated modular arithmetic, large ciphertexts, or multiple communication rounds. Model compression and compact architectures reduce the baseline cost of neural computation [11], [12], but they do not eliminate the overhead of secure computation. The design problem is therefore to minimize total system energy while maintaining acceptable privacy, accuracy, and latency.

Cryptographic primitives have distinct energy profiles. Homomorphic encryption can perform certain operations without interaction, but ciphertext expansion and computational complexity can be high. Secret sharing requires communication among parties but can be efficient for linear operations. Garbled circuits and oblivious transfer may be suitable for non-linear functions but can involve significant bandwidth. Secure multi-party computation frameworks often combine these techniques to balance cost across parties [8], [9], [10]. In resource-constrained settings, the choice of protocol must consider not only the device's own energy budget but also the energy consumed by gateways and servers, because the overall system may be powered by different stakeholders with different incentives.

Communication energy is particularly important when devices rely on wireless links with variable quality. Federated learning and secure aggregation reduce the need to transmit raw data, but they require multiple rounds of model updates [13], [14]. Each round consumes radio energy, and unreliable connections can cause retransmissions that further increase consumption. Edge gateways can aggregate updates locally, but this shifts energy use to a shared infrastructure that may also be constrained. Energy-aware scheduling, adaptive participation, and compressed update representations can help, but they must be integrated with privacy guarantees. For example, compressing updates too aggressively may weaken differential privacy or make secure aggregation less effective. Thus, energy efficiency is not a separate optimization layer but a constraint that shapes the entire privacy-preserving pipeline.

5. Privacy-Preserving Computation Strategies

Homomorphic encryption enables computation over encrypted data and is attractive for scenarios where devices can encrypt inputs and servers can process them without direct access to plaintext. It is especially useful for linear operations such as matrix multiplication, which dominate many neural network layers [5], [6]. However, non-linear activation functions remain challenging, and ciphertext expansion increases storage and bandwidth requirements. In resource-constrained internet-of-things systems, homomorphic encryption is often best suited to gateway-assisted or server-assisted workflows where the device performs only lightweight encryption and decryption.

Secure multi-party computation, including secret sharing and garbled circuits, allows multiple parties to jointly evaluate a function without revealing their inputs. This paradigm is well suited to private inference when clients and model owners do not trust each other [1], [2], [3], [7], [8], [9], [10]. Secret-sharing protocols can be efficient for linear layers, while garbled circuits or oblivious transfer handle comparisons and activation functions. The main drawback is communication overhead, which can be problematic for battery-powered devices. Hybrid protocols attempt to mitigate this by assigning different operations to the most efficient primitive. The resulting system is often faster than a single-technique approach, but it is also more complex to implement, audit, and maintain.

Differential privacy provides a statistical guarantee that the inclusion or exclusion of any single data record has limited effect on the output. It can be applied to model training, inference, or aggregation, and it is often combined with federated learning to protect individual contributions [15], [18]. Differential privacy is valuable because it offers quantifiable protection even when cryptographic assumptions fail, but it introduces noise that can reduce accuracy. In internet-of-things settings, the privacy budget must be allocated across many devices and repeated queries, making long-term privacy management a governance challenge as much as a technical one. Federated learning, meanwhile, keeps raw data on devices and shares model updates instead [13], [14]. Secure aggregation can further protect updates from the server, but it increases communication and coordination costs. These strategies are complementary rather than mutually exclusive, and their combination determines both privacy and energy outcomes.

6. Structural Trade-offs and Optimization for Energy-Efficient Private Inference

The central trade-off in private neural computation at the edge is among privacy, accuracy, latency, and energy. Stronger cryptographic protection often increases computation and communication, while aggressive compression can reduce energy but may harm accuracy or weaken privacy. A system designer must therefore decide which operations occur on the device, which occur at the gateway, and which occur in the cloud. These decisions are structural because they determine trust boundaries, failure modes, and the distribution of costs across stakeholders. They also affect fairness, since devices with fewer resources may be unable to participate in protocols that require substantial computation or bandwidth [9], [10], [11], [12].

Activation functions are a particularly important source of overhead in secure inference. Non-linear operations such as rectified linear units are efficient in plaintext computation but expensive in secure two-party or multi-party protocols because they require comparisons or oblivious selection. Recent work has explored ReLU-reduced neural network designs for two-party computation based private inference, showing that architectural changes can lower cryptographic cost without abandoning privacy goals [16]. This line of research is significant for resource-constrained internet-of-things systems because it treats the neural network architecture and the secure protocol as co-designed artifacts. Instead of optimizing the protocol alone, it reduces the number of expensive non-linear operations that the protocol must handle. Such co-design can produce substantial energy savings, but it may require retraining, changes in model accuracy, and careful validation across diverse data distributions.

Hybrid secure computation frameworks also offer a structural response to energy constraints. By assigning linear operations to secret sharing and non-linear operations to garbled circuits or oblivious transfer, they can reduce the total cost of private inference [7], [8], [10]. Secure offloading extends this logic by moving expensive operations from the device to a gateway or

edge server, provided that the trust model permits it. The trade-off is that offloading increases network dependence and may expose metadata such as timing, frequency, and model choice. Hardware acceleration, including trusted execution environments, cryptographic coprocessors, and specialized neural accelerators, can further improve energy efficiency, but it introduces supply-chain and attestation concerns. The optimal configuration is therefore context-dependent and must be evaluated against deployment goals, regulatory requirements, and the capabilities of the target device population [11], [12], [16].

7. Deployment, Governance, and Policy Implications

Deploying privacy-preserving neural networks in internet-of-things ecosystems requires governance arrangements that match the technical architecture. Data protection regulations often require data minimization, purpose limitation, and accountability, but secure computation can complicate the question of who is a data controller and who is a processor. If a gateway performs encrypted inference and never sees plaintext, its legal role may differ from that of a cloud provider that decrypts intermediate results. These distinctions must be reflected in contracts, audit trails, and incident response plans. Without such governance, technical privacy guarantees may be undermined by unclear responsibility or inconsistent operational practices [17], [18].

Policy frameworks also need to address cross-border data flows, sector-specific rules, and public procurement. In healthcare, for example, private inference on wearable devices may fall under medical device regulation, privacy law, and cybersecurity standards. In industrial settings, proprietary models and trade secrets may be as sensitive as personal data. Secure computation can support compliance by reducing the need to transfer raw data, but it does not automatically resolve legal obligations. Regulators and standards bodies should therefore encourage transparency about protocol assumptions, adversarial models, and residual leakage. Auditable design documentation can help third parties assess whether a system's privacy claims are meaningful in practice [19].

Fairness is another governance concern. If privacy-preserving inference requires more computation or bandwidth, devices in low-resource or rural settings may be excluded or forced to accept weaker privacy protections. This creates a two-tier system in which wealthier users and better-connected regions receive stronger privacy and better performance. Policy interventions such as public edge infrastructure, open standards, and energy-aware procurement can reduce these disparities. Differential privacy and federated learning can help, but they must be implemented with attention to group fairness and representation. If models are trained predominantly on data from well-resourced populations, private inference may still produce biased outcomes. Governance must therefore connect privacy protection to broader questions of distributive justice and inclusive innovation [20].

8. Robustness, Fairness, and Sustainability

Robustness is essential because secure computation protects confidentiality but does not guarantee integrity or availability. An adversary may attempt to poison training data, manipulate model updates, or induce denial of service by exhausting communication or battery resources. In federated settings, malicious clients can submit corrupted updates, and secure aggregation may hide their identities, making detection more difficult. Differential privacy can limit the impact of individual records, but it does not prevent coordinated poisoning. Robust aggregation, anomaly detection, and reputation systems can help, but they

add computation and communication overhead. The design challenge is to achieve robustness without undermining energy efficiency or privacy guarantees [20].

Fairness and sustainability are closely linked in edge intelligence. Fairness requires that models perform equitably across demographic groups, device types, and environmental conditions. Sustainability requires that the energy and material costs of devices, batteries, and network infrastructure be considered over the full lifecycle. Privacy-preserving computation can reduce data transmission and therefore lower some energy costs, but cryptographic overhead may increase others. The net environmental effect depends on the deployment context, including the energy mix of the grid, the lifetime of devices, and the frequency of inference. A systems perspective must therefore evaluate not only operational energy but also embodied energy, e-waste, and maintenance burdens [11], [12], [17].

Sustainable deployment also depends on governance that supports long-term maintenance and interoperability. If privacy-preserving protocols are proprietary or poorly documented, devices may become obsolete when vendors withdraw support, leading to premature replacement and additional e-waste. Open standards, modular architectures, and transparent security updates can extend device lifetimes and reduce environmental harm. At the same time, regulators should avoid mandating specific cryptographic techniques that may become outdated. Instead, they can require demonstrable privacy properties, independent audits, and clear disclosure of trade-offs. This approach allows innovation while maintaining accountability. In this sense, energy-efficient private neural computation is not only a technical achievement but also an institutional commitment to responsible infrastructure [16], [17].

9. Conclusion

This paper has examined energy-efficient privacy-preserving neural networks for resource-constrained internet-of-things devices using secure computation. The analysis shows that privacy and energy efficiency are deeply entangled with system architecture, protocol choice, model design, and governance. Secure multi-party computation, homomorphic encryption, differential privacy, and federated learning each offer distinct advantages, but none is a universal solution. Practical deployments will likely rely on hybrid protocols, tiered computation across devices and gateways, model compression, and architectural changes that reduce expensive non-linear operations. The co-design of neural networks and secure protocols, including ReLU-reduced approaches, illustrates how cryptographic cost can be lowered without abandoning privacy goals [16].

The paper also emphasizes that technical design cannot be separated from policy and sustainability. Trust assumptions must be transparent, accountability must be enforceable, and fairness must be considered across device populations and regions. Energy consumption must be evaluated at the system level, including communication, computation, and lifecycle impacts. Future research should develop standardized benchmarks for privacy, energy, latency, and fairness in edge environments; investigate adaptive protocols that respond to battery state and network conditions; and create governance mechanisms that support auditability without stifling innovation. If these challenges are addressed, secure neural computation can become a foundation for trustworthy, inclusive, and environmentally responsible edge intelligence.

References

1. Yao, A. C. (1982). Protocols for secure computations. In 23rd Annual Symposium on Foundations of Computer Science (pp. 160–164). IEEE.
<https://doi.org/10.1109/SFCS.1982.38>
2. Goldreich, O., Micali, S., & Wigderson, A. (1987). How to play any mental game. In Proceedings of the 19th Annual ACM Symposium on Theory of Computing (pp. 218–229). ACM. <https://doi.org/10.1145/28395.28420>
3. Beaver, D. (1991). Efficient multiparty protocols using circuit randomization. In J. Feigenbaum (Ed.), *Advances in cryptology—CRYPTO '91* (pp. 420–432). Springer.
https://doi.org/10.1007/3-540-46766-1_34
4. Canetti, R. (2001). Universally composable security: A new paradigm for cryptographic protocols. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science (pp. 136–145). IEEE. <https://doi.org/10.1109/SFCS.2001.959888>
5. Damgård, I., Pastro, V., Smart, N., & Zakarias, S. (2012). Multiparty computation from somewhat homomorphic encryption. In R. Safavi-Naini & R. Canetti (Eds.), *Advances in cryptology—CRYPTO 2012* (pp. 643–662). Springer. https://doi.org/10.1007/978-3-642-32009-5_38
6. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on Machine Learning (pp. 201–210). PMLR. <https://proceedings.mlr.press/v48/gilad-bachrach16.html>
7. Liu, J., Juuti, M., Lu, Y., & Asokan, N. (2017). Oblivious neural network predictions via MiniONN transformations. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 619–631). ACM.
<https://doi.org/10.1145/3133956.3134056>
8. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. In 2017 IEEE Symposium on Security and Privacy (pp. 19–38). IEEE.
<https://doi.org/10.1109/SP.2017.12>
9. Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A. (2018). Gazelle: A low latency framework for secure neural network inference. In 27th USENIX Security Symposium (pp. 1651–1669). USENIX Association.
<https://www.usenix.org/conference/usenixsecurity18/presentation/juvekar>
10. Riazi, M. S., Weinert, C., Tkachenko, O., Songhori, E. M., Schneider, T., & Koushanfar, F. (2018). Chameleon: A hybrid secure computation framework for machine learning applications. In Proceedings of the 2018 on Asia Conference on Computer and Communications Security (pp. 707–721). ACM.
<https://doi.org/10.1145/3196494.3196522>
11. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In 4th International Conference on Learning Representations. <https://arxiv.org/abs/1510.00149>
12. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. <https://arxiv.org/abs/1704.04861>

13. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR.
<https://proceedings.mlr.press/v54/mcmahan17a.html>
14. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175–1191). ACM.
<https://doi.org/10.1145/3133956.3133982>
15. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318). ACM.
<https://doi.org/10.1145/2976749.2978318>
16. Peng, Hongwu, et al. "Rrnet: Towards relu-reduced neural network for two-party computation based private inference." *arXiv preprint arXiv:2302.02292* (2023).
17. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
18. Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, languages and programming* (pp. 1–12). Springer.
https://doi.org/10.1007/11787006_1
19. Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., & Ludwig, H. (2019). HybridAlpha: An efficient approach for privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security* (pp. 13–23). ACM.
<https://doi.org/10.1145/3338501.3357371>
20. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations*. <https://arxiv.org/abs/1610.05755>