

Energy-Efficient Secure Inference for Neural Networks on Edge Devices

Mathijs van Haeften

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.

vanhaefte1994@oregonstate.edu

Abhimanyu Lalla

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

lalla1976@unh.edu

Leopoldo Puccini

School of Computing, Clemson University, Clemson, SC, USA.

leopoldo462@clemson.edu

Kirstin Leen

Department of Computer Science, George Mason University, Fairfax, VA, USA.

kirstinleen247@gmu.edu

Abstract

Secure inference for neural networks on edge devices sits at the intersection of cryptographic protection, constrained computation, and operational sustainability. Edge devices increasingly perform or assist with inference because they reduce latency, preserve local context, and lessen dependence on centralized cloud infrastructure. Yet privacy-preserving mechanisms such as homomorphic encryption, secure multi-party computation, and trusted execution environments impose substantial energy, memory, and communication costs. This paper examines the system-level trade-offs that shape energy-efficient secure inference, with attention to architecture, hardware heterogeneity, deployment governance, robustness, fairness, and policy. It argues that secure inference should not be treated as a standalone cryptographic layer but as a socio-technical infrastructure problem. The analysis considers how model compression, quantization, protocol design, and edge-cloud partitioning interact with security guarantees and energy budgets. It also explores how measurement practices, supply chains, standards, and regulatory expectations influence adoption. The paper concludes that durable progress requires co-design across algorithms, protocols, hardware, and institutions, and that energy efficiency must be evaluated alongside privacy, accuracy, resilience, and equity rather than as a substitute for them.

Keywords

energy-efficient secure inference, edge computing, privacy-preserving machine learning, two-party computation, homomorphic encryption, system architecture, sustainability, governance.

1. Introduction

Neural network inference is moving rapidly from centralized data centers toward edge devices, including smartphones, wearables, cameras, vehicles, and industrial sensors. This shift is driven by latency requirements, bandwidth constraints, intermittent connectivity, and the

desire to keep sensitive data closer to its source. Edge computing provides a conceptual and infrastructural foundation for this movement by distributing computation and storage across the network continuum rather than concentrating everything in remote clouds [1]. However, edge inference also creates new security and privacy challenges because devices operate in physically accessible, heterogeneous, and often untrusted environments. The result is a demanding design space in which accuracy, latency, energy, privacy, and maintainability must be balanced simultaneously.

Secure inference techniques aim to protect model parameters, user inputs, or both during computation. Cryptographic approaches such as secure multi-party computation and homomorphic encryption offer strong formal guarantees, but they are computationally expensive and often unsuitable for battery-powered devices without careful optimization [8][9][10]. Trusted execution environments provide another route, yet they depend on hardware vendor trust, attestation, and side-channel resistance. In practice, secure inference on edge devices is rarely a single mechanism. It is a composition of model optimizations, protocol choices, hardware capabilities, orchestration logic, and governance arrangements. Understanding this composition is essential because a technique that appears efficient in one deployment context may be unacceptable in another.

This paper argues that energy-efficient secure inference must be approached as a system-level and socio-technical problem. It is not enough to reduce cryptographic operations if communication overhead, memory movement, or thermal throttling dominates the energy profile. It is also not enough to achieve low energy consumption if the resulting system is brittle, unfair, or impossible to audit. The discussion therefore spans architectural patterns, energy accounting, hardware heterogeneity, robustness, fairness, deployment operations, and policy. The central claim is that sustainable secure inference requires co-design across algorithmic, infrastructural, and institutional layers, with evaluation practices that reflect real deployment constraints rather than isolated benchmark victories.

2. Background and Motivation

The growth of edge intelligence is supported by advances in model efficiency and mobile hardware. Surveys of deep learning in mobile and wireless networking show that inference workloads are increasingly distributed across devices, gateways, and cloud resources [3]. Lightweight neural network families such as MobileNets and MobileNetV2 demonstrate that architectural choices can dramatically reduce computation while preserving acceptable accuracy for many vision tasks [4][5]. Compression techniques such as pruning, quantization, and Huffman coding further reduce model size and memory bandwidth, which are often more limiting than raw arithmetic throughput on edge devices [6]. Integer-only quantization has also matured to the point where efficient inference can be performed without floating-point support, a critical enabler for low-power processors [7].

Parallel to these efficiency advances, privacy-preserving inference has developed a rich set of cryptographic protocols. Early work on secure computation established the theoretical foundations for evaluating functions without revealing inputs [8][9]. Fully homomorphic encryption demonstrated that computation over encrypted data is possible, though initially at prohibitive cost [10]. Subsequent systems such as CryptoNets, SecureML, MiniONN, and Gazelle translated these ideas into practical neural network inference, using techniques such as oblivious transfer, secret sharing, and garbled circuits [11][12][13][14]. More recent research has focused on reducing the cost of non-linear activation functions and other expensive operations in two-party settings [15]. Frameworks such as CrypTen, ABY3, and

CrypTFlow2 have further improved usability and performance for secure machine learning, but they still impose significant overhead relative to plaintext inference [16][17][18].

At the same time, federated learning has emerged as a complementary approach that keeps raw data on devices while sharing model updates [19]. Federated learning reduces some privacy risks but does not eliminate them, and it introduces its own communication and energy costs. Sustainability concerns have also entered the mainstream, with studies of energy and policy considerations in deep learning and the Green AI movement arguing for broader efficiency metrics [20][21]. Industrial analyses of inference at the edge show that deployment scale, request patterns, and hardware utilization strongly affect energy outcomes [22]. The motivation for this paper is therefore to connect these strands: cryptographic security, edge efficiency, operational sustainability, and governance must be considered together if secure inference is to become a practical default rather than a niche capability.

3. Threat Models and Security Objectives for Edge Inference

Threat models for edge inference vary widely. In one common scenario, a user sends an input to a remote server that holds a proprietary model. The user may want to protect the input, while the model owner may want to protect the model. In another scenario, inference runs entirely on a user device, and the primary concern is protecting the model from extraction or tampering. In a third scenario, computation is split across a device, a gateway, and a cloud service, creating multiple trust boundaries and potential leakage points. Each scenario implies different security objectives, including confidentiality of inputs, confidentiality of weights, integrity of computation, availability under attack, and auditability of outcomes.

Cryptographic secure inference primarily addresses confidentiality and integrity. Homomorphic encryption allows computation on encrypted inputs, while secure multi-party computation distributes trust among participants so that no single party sees the full data. These guarantees are attractive, but they are not absolute. Side channels, metadata leakage, access patterns, and implementation flaws can undermine theoretical protections. Trusted execution environments can provide strong isolation, but they rely on hardware root of trust and are vulnerable to microarchitectural attacks. Consequently, threat modeling must include not only the cryptographic protocol but also the operating system, firmware, network stack, and physical environment.

Security objectives also interact with privacy regulation and organizational accountability. A system that protects input data but leaks model behavior through repeated queries may still violate expectations of privacy or intellectual property. A system that protects model weights but produces biased or unexplainable decisions may fail fairness and due process requirements. For edge deployments in healthcare, transportation, or public infrastructure, availability and safety may be as important as confidentiality. Thus, secure inference should be evaluated against a portfolio of objectives rather than a single cryptographic metric. Governance mechanisms such as logging, attestation, access control, and independent audit are necessary to translate technical guarantees into institutional trust.

4. System Architectures for Energy-Efficient Secure Inference

Architectural choices determine the energy profile of secure inference more than any single algorithm. A fully local architecture avoids network transmission but requires the edge device to perform both plaintext inference and cryptographic protection. This may be feasible for small models and trusted execution environments, but it becomes difficult for large models and strong cryptographic protocols. A fully remote architecture shifts computation to more

capable servers, but it increases communication energy and exposes data to network and server-side risks. Split inference partitions the model between device and server, often placing early layers on the device and later layers in the cloud. Secure split inference can protect intermediate representations, but it may still reveal information through the partition point and timing.

Two-party computation is a natural fit for client-server secure inference because it can distribute trust between the user and the model provider. However, two-party protocols often require multiple rounds of interaction, large communication volumes, and expensive non-linear operations. Recent work on reducing non-linear activation costs illustrates how protocol-aware model design can improve practicality [15]. Multi-party computation extends this to more than two parties, which can reduce trust assumptions but increases coordination overhead. Homomorphic encryption can support non-interactive computation, but bootstrapping and large ciphertext expansion remain challenging for edge devices. Hybrid architectures that combine trusted execution environments with cryptographic protocols attempt to balance performance and trust, though they inherit the limitations of both.

Hardware acceleration is another architectural dimension. Graphics processing units, digital signal processors, neural processing units, and field-programmable gate arrays can accelerate plaintext inference, but cryptographic operations may not map well to these units. Specialized accelerators for homomorphic encryption and secure computation are emerging, yet they are not yet common in edge devices. Batching, pipelining, and quantization can improve throughput and reduce energy per inference, but they may alter security assumptions or introduce new leakage. The most energy-efficient architecture is therefore context-dependent: it depends on model size, privacy requirements, network conditions, hardware capabilities, and the cost of trust. A system-level approach must treat these factors as coupled rather than independent.

5. Energy Accounting, Hardware Heterogeneity, and Sustainability

Energy accounting for secure inference is notoriously difficult. Runtime is a poor proxy for energy because different operations draw different amounts of power. Cryptographic computations may be energy-intensive in bursts, while communication radios may consume significant energy even when idle. Memory movement, cache behavior, and thermal throttling further complicate measurement. On edge devices, battery capacity and thermal design power constrain sustained performance, so a protocol that is fast in a short benchmark may be unsustainable over hours of use. Accurate energy accounting requires instrumentation at the hardware, operating system, and application levels, along with workload models that reflect real request distributions.

Hardware heterogeneity compounds the problem. A smartphone, a microcontroller, an automotive computer, and an industrial gateway have vastly different capabilities and energy characteristics. A secure inference protocol optimized for one may be inefficient or impossible on another. Model compression and quantization can reduce energy on some devices but increase overhead on others due to conversion costs or reduced parallelism. Trusted execution environments may offer energy savings by avoiding heavy cryptography, but they may also force data through secure memory regions that consume additional power. Sustainability therefore cannot be reduced to a single efficiency number; it must consider embodied carbon, device lifetime, replacement cycles, and the energy mix of the surrounding infrastructure.

From a sustainability perspective, edge inference can be beneficial when it avoids unnecessary data transmission and centralized cooling. However, it can also shift energy consumption to millions of distributed devices, making aggregate impact harder to monitor. If secure inference requires frequent protocol rounds or large ciphertext exchanges, the network energy may outweigh the savings from local computation. Lifecycle analysis should include manufacturing, deployment, operation, and disposal. A device that performs secure inference slowly may need a larger battery or more frequent charging, which affects user behavior and grid demand. These considerations suggest that energy efficiency must be evaluated across the entire service lifecycle rather than only at the inference step.

6. Privacy, Fairness, Robustness, and Governance

Privacy, fairness, and robustness are often treated as separate concerns, but they are deeply intertwined in secure edge inference. A privacy-preserving system may still produce discriminatory outcomes if the model was trained on biased data or if the secure protocol obscures the need for correction. Conversely, a fair model may leak sensitive attributes through side channels or model updates. Robustness is also relevant: adversarial inputs, model extraction attacks, and denial-of-service attempts can degrade both security and energy efficiency. A system that is secure under normal conditions but fails catastrophically under attack may be unacceptable in safety-critical settings.

Governance mechanisms are needed to manage these interactions. Audit trails can support accountability, but they must themselves be privacy-preserving and energy-efficient. Attestation can verify that a device is running approved software, yet it may exclude older or lower-cost devices, raising equity concerns. Differential privacy can provide formal privacy guarantees, but it may reduce accuracy and increase training or inference costs. Secure multi-party computation can protect inputs, but it may make debugging and error correction more difficult. Policymakers and system designers must therefore consider not only whether a technique is technically sound but also whether it can be operated transparently, contested, and improved over time.

Fairness also has an infrastructural dimension. If secure inference is only available on high-end devices, then privacy and security become privileges rather than rights. If edge deployments concentrate in wealthy regions or well-resourced institutions, the benefits of low-latency private inference may bypass marginalized communities. Governance should therefore include procurement standards, accessibility requirements, and support for interoperable open protocols. Energy efficiency should not be pursued by weakening security for users who cannot afford stronger protection. Instead, the goal should be to develop scalable and adaptable mechanisms that can provide robust protection across diverse device classes and deployment contexts.

7. Deployment, Operations, and Ecosystem Coordination

Deploying secure inference at scale requires operational capabilities that go beyond model development. Devices must be provisioned with keys, certificates, and secure software images. Updates must be delivered reliably without disrupting inference or compromising security. Monitoring must detect failures, attacks, and performance degradation without collecting excessive personal data. Orchestration must decide when to run locally, when to offload, and when to invoke secure computation. These decisions depend on network conditions, battery state, thermal limits, and user preferences. MLOps practices for edge AI must therefore be extended to include cryptographic key management, attestation, and energy-aware scheduling.

Ecosystem coordination is equally important. Hardware vendors, protocol designers, application developers, cloud providers, and regulators must align on interfaces and expectations. Standards for secure inference are still emerging, and fragmentation can increase costs and reduce trust. Interoperability between trusted execution environments, cryptographic libraries, and model formats would lower barriers to adoption. At the same time, premature standardization can lock in inefficient or insecure designs. A balance is needed between stability and experimentation. Open benchmarks and reference implementations can help, but they must capture energy, latency, accuracy, and security jointly rather than optimizing one dimension in isolation.

Economic and organizational factors also shape deployment. Secure inference may increase cloud costs, device costs, or operational complexity. Organizations may resist adoption if benefits are diffuse and costs are concentrated. Business models that align incentives are therefore necessary. For example, model providers may offer secure inference as a premium service, while device manufacturers may differentiate on privacy features. Insurance and liability regimes may reward demonstrable security. Public procurement can stimulate demand for energy-efficient secure inference in healthcare, public safety, and critical infrastructure. Without such alignment, technically superior solutions may remain underdeployed.

8. Evaluation Challenges and Comparative Perspectives

Evaluating secure inference is difficult because there is no single metric that captures all relevant qualities. Accuracy, latency, throughput, energy per inference, communication volume, memory footprint, and security level are all important, but they trade off against one another. A protocol that is fast on a powerful server may be unusable on a microcontroller. A model that is accurate on a benchmark may be fragile in the field. An energy measurement that ignores network transmission may understate the true cost. Comparative studies must therefore specify the deployment scenario, threat model, hardware platform, and workload distribution. Otherwise, results are not comparable and can mislead practitioners.

Cross-domain comparisons reveal different priorities. In healthcare, privacy and auditability may dominate, and energy efficiency may be secondary to correctness and regulatory compliance. In autonomous vehicles, latency and safety are paramount, and secure inference must operate under strict real-time constraints. In smart cities, scale and maintainability matter, and energy efficiency is tied to sustainability goals. In consumer devices, battery life and user experience are central, and security must be unobtrusive. These differences mean that a one-size-fits-all approach is unlikely to succeed. Instead, modular architectures and configurable protocols can allow the same underlying technology to be adapted to different domains.

Forward-looking evaluation should also consider adversarial and long-term conditions. How does energy consumption change under attack? How does performance degrade when devices are older or networks are congested? How does the system behave when keys are compromised or updates fail? How are fairness and privacy monitored over time? These questions require longitudinal studies and field deployments, not only laboratory benchmarks. They also require interdisciplinary collaboration among cryptographers, systems researchers, social scientists, and policymakers. The evaluation challenge is therefore not merely technical; it is methodological and institutional.

9. Policy, Standards, and Future Research Directions

Policy frameworks for edge AI are still evolving. Data protection regulations emphasize privacy, but they often do not specify how secure inference should be implemented. Cybersecurity standards focus on resilience, but they may not address energy efficiency. Sustainability policies promote green computing, but they rarely consider the energy cost of cryptographic protection. A more integrated approach would recognize that privacy, security, and sustainability are interdependent. Procurement rules, certification schemes, and reporting requirements could encourage energy-efficient secure inference without mandating specific technologies. They could also support transparency about trade-offs and enable independent verification.

Future research should pursue co-design across layers. At the algorithmic level, model architectures can be designed to reduce expensive secure operations, as illustrated by recent work on activation reduction [15]. At the protocol level, adaptive schemes can select different security mechanisms based on context, device capability, and energy availability. At the hardware level, accelerators for both neural inference and cryptographic primitives can reduce energy per operation. At the system level, schedulers can decide when to compute locally, when to offload, and when to defer tasks to save energy. At the institutional level, standards and governance can create incentives for adoption and accountability.

Several open questions remain. How can energy efficiency be measured consistently across heterogeneous devices and networks? How can secure inference be made accessible to low-resource users without compromising protection? How can fairness be audited when data and computation are distributed? How can systems recover gracefully from attacks or failures? How can supply chains and lifecycle impacts be included in sustainability assessments? Addressing these questions will require sustained investment in research infrastructure, open datasets, and multidisciplinary collaboration. The goal should not be to optimize one dimension at the expense of others, but to build systems that are secure, efficient, fair, and durable enough for long-term deployment.

10. Conclusion

Energy-efficient secure inference on edge devices is a complex systems problem that cannot be solved by cryptography alone. It requires architectures that balance local computation, offloading, hardware acceleration, and trust assumptions. It requires energy accounting that captures computation, communication, memory, and lifecycle effects. It requires governance that addresses privacy, fairness, robustness, and accountability. It also requires policy and standards that align incentives across vendors, operators, regulators, and users. The most promising path forward is co-design: model architectures that respect protocol costs, protocols that adapt to device conditions, hardware that supports both neural and cryptographic workloads, and institutions that reward sustainable and equitable deployment. Secure inference should be treated not as a feature added at the end of development but as an infrastructural commitment that shapes how edge intelligence evolves. By evaluating privacy, energy, accuracy, and resilience together, researchers and practitioners can advance systems that are not only technically sophisticated but also socially and environmentally responsible.

References

1. Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30-39.
<https://doi.org/10.1109/MC.2017.9>

2. Alrawais, A., Alhothaily, A., Hu, C., & Cheng, X. (2017). Fog computing for the Internet of Things: Security and privacy issues. *IEEE Internet Computing*, 21(2), 34-42.
<https://doi.org/10.1109/MIC.2017.37>
3. Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(3), 2224-2287.
<https://doi.org/10.1109/COMST.2019.2904897>
4. 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*. <https://arxiv.org/abs/1704.04861>
5. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510-4520.
<https://doi.org/10.1109/CVPR.2018.00474>
6. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *International Conference on Learning Representations*.
7. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2704-2713. <https://doi.org/10.1109/CVPR.2018.00286>
8. Yao, A. C. C. (1986). How to generate and exchange secrets. *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, 162-167.
<https://doi.org/10.1109/SFCS.1986.25>
9. Goldreich, O., Micali, S., & Wigderson, A. (1987). How to play any mental game. *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*, 218-229.
<https://doi.org/10.1145/28395.28420>
10. Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 169-178.
<https://doi.org/10.1145/1536414.1536440>
11. 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. *International Conference on Machine Learning*, 201-210.
12. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. *2017 IEEE Symposium on Security and Privacy*, 19-38.
<https://doi.org/10.1109/SP.2017.12>
13. Liu, J., Juuti, M., Lu, Y., & Asokan, N. (2017). Oblivious neural network predictions via MiniONN transformations. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 619-631.
<https://doi.org/10.1145/3133956.3134056>
14. Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A. (2018). GAZELLE: A low latency framework for secure neural network inference. *Proceedings of the 27th USENIX Security Symposium*, 1651-1669.

15. Peng, H., Zhou, S., Luo, Y., Xu, N., Duan, S., Ran, R., ... & Ding, C. (2023). Rrnet: Towards relu-reduced neural network for two-party computation based private inference. arXiv preprint arXiv:2302.02292.
16. Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., & van der Maaten, L. (2021). CrypTen: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34.
17. Mohassel, P., & Rindal, P. (2018). ABY3: A mixed protocol framework for machine learning. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 35-52. <https://doi.org/10.1145/3243734.3243760>
18. Rathee, D., Rathee, M., Kumar, N., Chandran, P., Gupta, A., Rastogi, A., & Sharma, R. (2020). CrypTFlow2: Practical 2-party secure inference. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 325-342. <https://doi.org/10.1145/3372297.3417274>
19. 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.
20. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645-3650. <https://doi.org/10.18653/v1/P19-1355>
21. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54-63. <https://doi.org/10.1145/3381831>
22. Wu, C. J., Brooks, D., Chen, K., Chen, D., Choudhury, S., Dukhan, M., ... & Zhang, P. (2019). Machine learning at Facebook: Understanding inference at the edge. *2019 IEEE International Symposium on High Performance Computer Architecture*, 331-344. <https://doi.org/10.1109/HPCA.2019.00048>