Model Compression Techniques for Privacy-Preserving Neural Network Inference

Jafet Serna
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
jafet1986@unh.edu

Willem Suijker
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
willem.suijker@missouri.edu

Abstract

Privacy-preserving neural network inference has emerged as a critical area of socio-technical infrastructure because machine learning systems increasingly mediate decisions in medicine, finance, public administration, and personal communication. The central challenge is not only to preserve predictive utility but also to protect input data, model parameters, and intermediate representations from adversarial inference. Cryptographic inference, trusted execution, federated learning, and differential privacy provide important safeguards, yet they impose substantial computational, communication, memory, and orchestration costs. Model compression techniques, including pruning, quantization, distillation, sparsity induction, and activation replacement, are often framed as efficiency optimizations. This paper argues that within privacy-preserving inference they function as architectural and governance mechanisms that reshape trust boundaries, deployment viability, robustness, fairness, and sustainability. The analysis examines how compression interacts with secure multiparty computation, homomorphic encryption, trusted execution environments, federated pipelines, and differential privacy mechanisms. It also considers the structural trade-offs between latency, accuracy, privacy leakage, auditability, and equity. Particular attention is given to nonlinear activation replacement, hardware-aware quantization, distributed model updates, and the policy implications of compressed private inference. The paper concludes that compression should be treated as a first-class design variable in privacy-preserving machine learning, requiring interdisciplinary evaluation across security, systems, law, and institutional accountability. Without such evaluation, compressed private inference may achieve short-term efficiency while introducing hidden vulnerabilities, inequitable performance, and opaque operational risks.

Keywords

privacy-preserving inference; model compression; secure multiparty computation; homomorphic encryption; federated learning; differential privacy; trustworthy machine learning; systems governance.

1. Introduction

The deployment of neural network inference in sensitive domains has transformed privacy from a compliance concern into a systems design constraint. Models trained on personal, medical, financial, or civic data can inadvertently reveal information about their training examples or about the inputs they process. Membership inference attacks demonstrate that an

adversary can determine whether a particular record was used during training, while model inversion attacks can reconstruct sensitive attributes from model outputs [1], [2]. Broader security analyses further show that machine learning pipelines exhibit attack surfaces spanning data collection, training, deployment, and feedback loops [3]. Consequently, privacy-preserving inference is not merely a cryptographic problem but a socio-technical infrastructure problem involving trust, accountability, and operational governance.

Cryptographic and trusted-computing approaches have been developed to enable inference without exposing raw inputs or model weights. Secure multiparty computation allows multiple parties to evaluate a function while keeping their respective inputs private, and systems such as SecureML and CryptoNets illustrate the feasibility of private neural network evaluation [4], [5]. Low-latency frameworks such as GAZELLE further demonstrate that secure inference can be optimized for practical deployment scenarios [6]. Nevertheless, these approaches often require heavy communication, large memory footprints, or specialized hardware, especially when nonlinear operations and deep architectures are involved. The resulting overhead can exclude smaller organizations, mobile clients, and public-interest applications from adopting strong privacy protections.

Model compression has therefore become a central response to the cost of private inference. Pruning removes redundant parameters, quantization reduces numerical precision, and coding schemes compress parameter storage and transmission [7]. Quantization-aware training and integer-only inference pipelines can significantly reduce computation while maintaining acceptable accuracy [8]. Recent work on automatic activation replacement shows that nonlinear components can be restructured to accelerate private network inference without abandoning cryptographic protections [9]. These techniques are usually presented as neutral efficiency tools, but in privacy-preserving settings they alter the geometry of trust, the distribution of computational responsibility, and the visibility of model behavior.

This paper advances the argument that model compression for privacy-preserving inference should be understood as a system-level design practice with security, fairness, sustainability, and policy consequences. The discussion proceeds through foundations of private inference, compression as an enabler, architectural trade-offs, robustness and fairness, governance, deployment infrastructures, case comparisons, and open challenges. The central claim is that compression cannot be evaluated only by accuracy and latency; it must also be assessed by its effects on leakage, auditability, inclusion, and long-term institutional trust.

2. Foundations of Privacy-Preserving Neural Network Inference

Privacy-preserving inference begins with a threat model. In some scenarios, the client wishes to protect its input from the model owner; in others, the model owner wishes to protect proprietary weights from the client; in still others, both parties require mutual protection. Membership inference and model inversion attacks show that even black-box access can leak sensitive information [1], [2]. Security and privacy analyses in machine learning further reveal that defenses often trade off against utility, interpretability, and deployability [3]. Therefore, any compression method must be evaluated against the specific adversary model it is expected to resist rather than against a generic notion of privacy.

The primary technical families for private inference include secure multiparty computation, homomorphic encryption, and trusted execution environments. SecureML and CryptoNets established early templates for encrypted or secret-shared neural computation [4], [5]. GAZELLE demonstrated that careful protocol design and optimized linear algebra can reduce

latency for secure inference [6]. However, these approaches remain sensitive to the depth of the network, the number of nonlinear operations, and the volume of intermediate data exchanged. Compression can reduce these burdens, but it can also introduce approximation errors that interact unpredictably with cryptographic protocols.

At the infrastructure level, private inference requires coordination among clients, servers, key managers, hardware accelerators, and auditors. Each component creates potential bottlenecks and trust assumptions. Trusted execution environments can simplify some cryptographic overhead, but they introduce hardware vendor dependencies and attestation requirements. Pure cryptographic methods avoid such dependencies but may demand extensive communication. Compression changes this landscape by lowering payload sizes, reducing memory pressure, and enabling more operations to occur at the edge. Yet it also complicates verification because compressed models may be harder to inspect, reproduce, or certify.

3. Model Compression as a Systems Enabler

Pruning, quantization, and coding are the classical pillars of model compression. Deep compression showed that removing redundant connections, reducing precision, and applying entropy coding can dramatically shrink neural networks [7]. In private inference, these reductions matter not only for storage and energy but also for the number of cryptographic operations and the size of encrypted messages. A smaller model can make secure evaluation feasible on devices that would otherwise be excluded. However, sparsity can create irregular computation patterns that are inefficient on some hardware and difficult to align with cryptographic primitives.

Quantization translates continuous or high-precision values into lower-precision representations. Quantization-aware training and integer-arithmetic inference allow models to run efficiently on commodity hardware while preserving accuracy [8]. For privacy-preserving inference, quantization can reduce the cost of secure multiplication, secret sharing, and homomorphic operations. Yet lower precision may weaken privacy guarantees if noise added for differential privacy is quantized away, or if quantization leaks information through timing and power side channels. Thus, quantization is not a purely computational choice; it is a privacy-relevant transformation.

Activation replacement illustrates the deep entanglement of compression and private inference. Nonlinear functions are often expensive in cryptographic protocols, and replacing or approximating them can substantially accelerate secure evaluation [9]. Knowledge distillation can transfer the behavior of a large model into a smaller student model, reducing inference cost while preserving approximate functionality [10]. Sparsity and lottery-ticket approaches suggest that trainable subnetworks can retain performance with fewer parameters [11]. These methods enable private inference at scale, but they also create new questions about fidelity, robustness, and whether the compressed model remains a faithful proxy for the original system.

Compression therefore functions as a systems enabler rather than a mere optimization. It determines which devices can participate, which protocols are economically viable, and which privacy guarantees remain meaningful after approximation. A compressed private model may be faster and cheaper, but if it excludes certain populations, amplifies errors for underrepresented groups, or resists external audit, then its efficiency gains are socially incomplete. The next section examines how these tensions manifest in concrete inference architectures.

4. Architectural Trade-Offs and Inference Pipelines

Private inference pipelines distribute computation across clients, servers, accelerators, and cryptographic engines. In secure multiparty computation, each party holds secret shares and collaborates on operations without revealing inputs. In homomorphic encryption, computation occurs directly on ciphertexts, but certain operations remain costly. In trusted execution environments, computation occurs inside an isolated enclave, shifting trust toward hardware attestation. Compression affects each architecture differently: it may reduce communication in multiparty settings, lower ciphertext expansion in homomorphic settings, or shrink memory footprints in enclave-based settings. These are structural trade-offs, not interchangeable performance gains.

Distributed and federated inference adds another layer of complexity. Gradient leakage attacks show that shared updates can reveal sensitive training data if protections are inadequate [12]. Differential privacy provides formal noise-based protections, but compression can interact with noise injection in ways that either strengthen or weaken the guarantee [13]. Federated learning enables decentralized training and inference across clients, yet communication efficiency remains a major constraint [14]. Broader surveys of federated learning identify privacy, heterogeneity, and system design as persistent open problems [15]. Compression can alleviate communication costs, but it can also introduce bias into aggregation and make adversarial contributions harder to detect.

Hardware and software co-design further shapes the design space. Efficient processing of deep neural networks depends on memory hierarchy, parallelism, data reuse, and operator support [16]. A compression technique that excels on a server-class accelerator may perform poorly on a mobile device or within a cryptographic coprocessor. Consequently, architecture decisions should be evaluated jointly with compression strategy, privacy protocol, and deployment environment. The most secure method is not necessarily the most sustainable, and the most efficient method is not necessarily the most equitable.

5. Privacy Guarantees, Robustness, and Fairness under Compression

Compression can affect robustness by changing the decision boundaries of a model. Pruning and quantization may remove rare but important features, making the model more vulnerable to adversarial examples or distribution shifts. In private inference, these vulnerabilities are compounded because the defender may not be able to inspect intermediate representations in plaintext. If a compressed model fails silently, clients may receive incorrect outputs without understanding whether the failure arises from compression, cryptographic approximation, or the underlying data distribution. Robustness therefore becomes a shared responsibility across model developers, protocol designers, and auditors.

Fairness is also implicated. Research on gender classification has shown that commercial systems can exhibit substantial accuracy disparities across intersectional groups [17]. Model cards provide a reporting mechanism for documenting intended use, performance, and limitations [18]. In compressed private inference, fairness reporting becomes more difficult because the deployed model may differ from the original due to pruning, quantization, distillation, or activation replacement. If performance disparities are not measured after compression, organizations may deploy a system that appears fair in development but behaves inequitably in production. Privacy protection does not automatically produce fairness; compression can obscure the very disparities that require remediation.

Differential privacy and federated learning offer important protections, but they do not eliminate these concerns. Differential privacy can limit the influence of any single record, yet it may reduce accuracy for minority subgroups if privacy budgets are allocated uniformly [13]. Federated learning can keep raw data local, but heterogeneous client data and uneven participation can produce models that perform poorly for underrepresented populations [14], [15]. Compression adds another layer of approximation that may interact with these effects. A socially robust private inference system must therefore evaluate privacy, robustness, and fairness together rather than treating them as separate compliance checks.

6. Governance, Policy, and Institutional Implications

Governance frameworks for private inference must address both technical and organizational accountability. Internal algorithmic auditing can help organizations trace how models are developed, deployed, monitored, and revised [19]. In compressed private inference, auditing requires documentation of compression choices, privacy parameters, hardware assumptions, and known failure modes. Without such records, oversight bodies cannot determine whether a system's privacy claims remain valid after deployment or whether compression has silently degraded protections. Governance thus depends on the same infrastructure that enables inference: logs, versioning, attestation, and reproducible evaluation.

Policy regimes increasingly demand data minimization, purpose limitation, and explainability. Privacy-preserving inference responds to these demands by reducing exposure of raw inputs and model weights. However, compression can complicate explainability because approximate models may be harder to interpret and because cryptographic protections may prevent direct inspection. Policymakers must therefore avoid treating privacy-enhancing technologies as automatic compliance solutions. A system can be cryptographically secure yet opaque, inequitable, or unsustainable. Effective regulation should require evidence across security, fairness, and environmental impact rather than focusing on a single technical guarantee.

Institutional trust also depends on procurement and standardization. If private inference tools are available only through proprietary hardware or closed cloud platforms, smaller public agencies and community organizations may be excluded. Open standards for compressed model formats, privacy accounting, and audit trails could reduce lock-in and improve comparability. At the same time, standards must be flexible enough to accommodate rapid advances in cryptography and model architecture. Governance should therefore be iterative, involving technologists, legal experts, civil society, and affected communities in ongoing assessment.

7. Deployment Infrastructures and Sustainability

The environmental cost of machine learning has become a central concern. Energy and policy considerations for deep learning show that training and deployment can produce significant carbon emissions [20]. Green AI argues for reporting efficiency alongside accuracy and for treating computational cost as a first-class research metric [21]. In privacy-preserving inference, these concerns are amplified because cryptographic operations and secure hardware can consume more energy than ordinary inference. Compression can reduce this burden, but it may also shift costs to edge devices, increase retraining frequency, or require specialized accelerators with embedded carbon footprints.

Sustainable deployment requires lifecycle analysis. A compressed model may use less energy per inference, but if it requires frequent recompression, retraining, or hardware replacement,

its total environmental impact may be higher. Secure multiparty computation may consume significant network bandwidth, while homomorphic encryption may demand large memory and computation. Federated learning can reduce centralized data transfer but may increase client-side computation and battery drain. These trade-offs cannot be resolved by a single metric; they require context-sensitive evaluation that includes energy sources, device lifetimes, and equity of access.

Infrastructure choices also affect resilience. Centralized private inference may be easier to monitor and secure, but it creates single points of failure. Decentralized approaches may be more resilient but harder to govern. Compression can support decentralization by reducing model size and communication requirements, enabling local inference on constrained devices. However, local deployment may reduce the ability to update models quickly in response to new threats. A sustainable private inference architecture must therefore balance central coordination with local autonomy, efficiency with adaptability, and privacy with accountability.

8. Case Illustrations and Cross-Domain Comparisons

In healthcare, private inference could allow hospitals to use remote diagnostic models without sharing patient data. Secure multiparty computation and homomorphic encryption can protect inputs, while compression reduces the cost of evaluating large imaging or genomic models [4], [5], [6]. However, medical models are often sensitive to small numerical changes, and aggressive quantization may reduce diagnostic accuracy for rare conditions. Governance must therefore include clinical validation after compression, not only cryptographic verification. The case illustrates that privacy and safety are intertwined: a private but inaccurate model is not clinically acceptable.

In finance, private inference may support fraud detection and credit scoring without exposing customer records. Federated learning can keep transaction data local, and differential privacy can limit disclosure [13], [14]. Compression can reduce communication costs across branches or mobile clients. Yet fairness risks are acute because credit decisions are legally regulated and historically discriminatory. If compression degrades performance for certain income groups or regions, the system may violate fair-lending principles. Auditors need access to post-compression performance metrics, not just privacy proofs. The financial case shows that governance must connect technical compression choices to legal non-discrimination requirements.

In public-sector applications, private inference may support benefits eligibility, identity verification, or civic analytics. These systems often involve vulnerable populations and require high transparency. Compression can enable deployment on low-cost devices, but it may also reduce explainability. If a compressed model denies a benefit or flags a case for review, affected individuals need a meaningful path to contest the decision. Model cards and algorithmic audits provide partial remedies [18], [19]. Cross-domain comparison reveals a common pattern: compression expands access and efficiency, but it also demands stronger evaluation, documentation, and redress mechanisms.

9. Open Challenges and Research Agenda

Several open challenges remain. First, privacy accounting under compression is immature. It is unclear how pruning, quantization, distillation, and activation replacement affect formal privacy guarantees in different protocols. Second, robust evaluation methodologies are needed to measure accuracy, fairness, and security after compression, especially when models are

deployed inside encrypted or attested environments. Third, hardware-software co-design must account for the joint constraints of cryptography, energy, and equity. Fourth, governance frameworks must evolve to audit compressed private systems without requiring disclosure of sensitive model or data details.

A forward-looking research agenda should treat compression as a cross-layer concern. It should develop standardized benchmarks that report privacy, accuracy, fairness, latency, energy, and auditability together. It should study how compression interacts with membership inference, model inversion, and gradient leakage [1], [2], [12]. It should examine how differential privacy and federated learning behave when models are compressed before, during, or after training [13], [14], [15]. It should also investigate participatory design methods that include affected communities in decisions about acceptable trade-offs. Without such work, private inference may become efficient but not trustworthy.

10. Conclusion

Model compression is often described as a way to make neural networks smaller and faster, but in privacy-preserving inference it performs a much broader role. It determines whether cryptographic protocols are practical, whether edge devices can participate, how much energy is consumed, and how easily systems can be audited. Techniques such as pruning, quantization, distillation, sparsity induction, and activation replacement can reduce the cost of private inference, but they also transform decision boundaries, trust assumptions, and accountability relationships. The required reference to automatic activation replacement illustrates how a technical optimization can directly enable faster private network inference while raising new questions about fidelity and verification [9].

This paper has argued that compression should be governed as a first-class design variable in privacy-preserving machine learning. Security evaluations must consider post-compression leakage and robustness. Fairness assessments must examine whether compressed models preserve equitable performance. Sustainability analyses must account for lifecycle energy and infrastructure dependencies. Governance frameworks must document compression choices, privacy parameters, and failure modes. If these concerns are treated as afterthoughts, private inference may achieve narrow efficiency goals while reproducing opacity, exclusion, and environmental harm. A more responsible trajectory requires interdisciplinary collaboration across systems, cryptography, law, policy, and the social sciences. The future of private inference depends not only on stronger protocols but also on how intelligently and inclusively compression is integrated into the broader socio-technical infrastructure of machine learning.

References

1. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3-18). IEEE. https://doi.org/10.1109/SP.2017.41

2. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333). Association for Computing Machinery. https://doi.org/10.1145/2810103.2813677

3. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. In 2018 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 399-414). IEEE. https://doi.org/10.1109/EuroSP.2018.00035

4. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 19-38). IEEE. https://doi.org/10.1109/SP.2017.12

5. 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.

6. 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.

7. 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. ICLR.

8. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neuralnetworks for efficient integer-arithmetic-only inference. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2704-2713). IEEE. https://doi.org/10.1109/CVPR.2018.00286

9. Peng, H., Huang, S., Zhou, T., Luo, Y., Wang, C., Wang, Z., ... & Ding, C. (2023, October). Autorep: Automatic relu replacement for fast private network inference. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5155-5165). IEEE.

10. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv. https://arxiv.org/abs/1503.02531

11. Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In International Conference on Learning Representations. ICLR.

12. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In Advances in Neural Information Processing Systems (pp. 14774-14784). Curran Associates.

13. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407. https://doi.org/10.1561/0400000042

14. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273-1282). PMLR.

15. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

16. Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329. https://doi.org/10.1109/JPROC.2017.2761740

17. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

19. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33-44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

20. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). Association for Computational Linguistics. 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