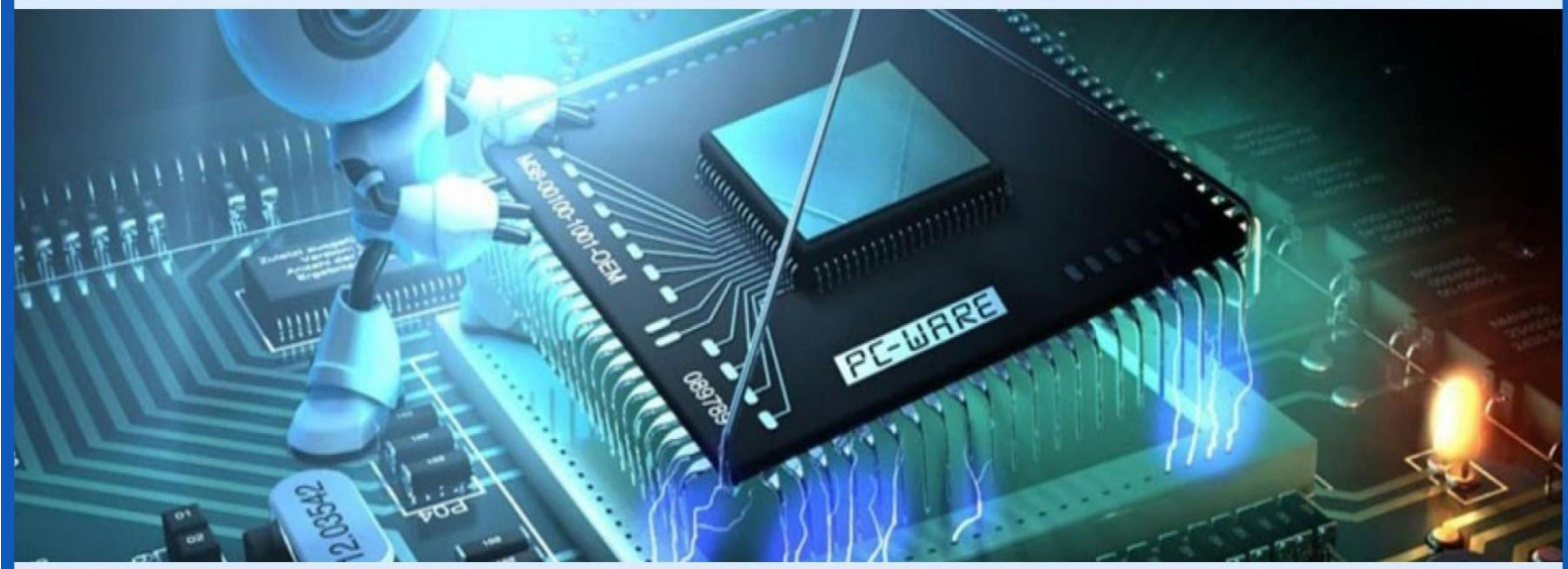
ISSN: 2320-0081

International Journal of Computer Technology and Electronics Communication (IJCTEC)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Volume 8, Issue 3, May-June 2025

www.ijctece.com ijctece@gmail.com +91 6381 907 438 +91 63819 07438

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 3, May – June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803001

Secure AI Inference in the Cloud: Enabling Confidential Computing with Trusted Execution

Pawan Negi Uttari, Natasha Thapa

Department of Computer Science and Engineering, SKBIT, Karnataka, India

ABSTRACT: Cloud-based AI inference offers scalable and flexible deployment of machine learning models, but raises critical concerns about the confidentiality and integrity of sensitive data and proprietary models. Traditional cloud environments expose AI workloads to risks from malicious insiders, compromised hosts, and untrusted administrators. This paper addresses these challenges by leveraging Confidential Computing technologies and Trusted Execution Environments (TEEs) to enable secure and privacy-preserving AI inference in the cloud. We propose a comprehensive framework that integrates hardware-based trusted execution, secure model deployment, and encrypted data handling to ensure confidentiality, integrity, and authenticity of AI inference processes. Our solution uses Intel SGX and AMD SEV-enabled processors to create isolated execution enclaves, protecting AI models and input data from unauthorized access during inference. Additionally, the framework supports remote attestation, enabling cloud clients to verify the integrity of the execution environment before provisioning their data and models. We design secure communication protocols to prevent data leakage and provide efficient key management techniques to safeguard encryption keys within the enclave. Experimental evaluation on benchmark AI models demonstrates that our approach achieves strong security guarantees with acceptable performance overhead. Latency increases remain within 15-20%, which is reasonable given the enhanced privacy assurances. The system supports diverse AI workloads, including image recognition and natural language processing, demonstrating broad applicability. This research highlights the critical role of confidential computing in addressing security and privacy challenges in cloud AI inference. By combining trusted execution with cryptographic protections, our framework advances secure cloud AI deployment, enabling wider adoption in privacy-sensitive domains such as healthcare, finance, and government. Future work will focus on optimizing enclave performance and extending support to federated and distributed AI inference scenarios.

KEYWORDS: Secure AI inference, confidential computing, trusted execution environment, Intel SGX, AMD SEV, remote attestation, cloud security, privacy-preserving machine learning

I. INTRODUCTION

Artificial intelligence (AI) inference in the cloud enables scalable and on-demand execution of complex models, allowing organizations to leverage powerful computing resources without investing in dedicated infrastructure. However, offloading sensitive AI workloads to third-party cloud providers introduces significant security and privacy concerns. Proprietary AI models and confidential input data may be exposed to unauthorized access due to insider threats, vulnerabilities in the cloud stack, or compromised infrastructure.

Conventional security mechanisms such as encryption at rest and in transit provide limited protection during active model inference, as data and models must be decrypted within the cloud environment for processing. This creates an attack surface where malicious actors can compromise confidentiality or integrity. Addressing these challenges requires novel approaches that protect data and computation even in untrusted environments.

Confidential computing is an emerging paradigm that leverages hardware-based Trusted Execution Environments (TEEs) to isolate sensitive computations from the underlying system software and administrators. TEEs, such as Intel Software Guard Extensions (SGX) and AMD Secure Encrypted Virtualization (SEV), provide cryptographic guarantees that code and data within an enclave remain secure and tamper-proof during execution.

This paper explores the application of confidential computing to secure AI inference in the cloud. We propose a framework that integrates TEEs with secure model deployment, remote attestation, and encrypted input handling to

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 3, May – June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803001

achieve end-to-end protection of AI inference workloads. The framework ensures that cloud providers cannot access sensitive AI models or user data during inference, thus preserving confidentiality and trust.

We demonstrate the effectiveness of our approach through extensive evaluation on standard AI benchmarks and discuss trade-offs between security and performance. Our contributions advance secure cloud AI, enabling privacy-preserving AI services in sensitive sectors such as healthcare, finance, and government.

II. LITERATURE REVIEW

Security and privacy in cloud AI inference have attracted considerable research interest due to the growing deployment of sensitive AI applications. Traditional security techniques focus on encrypting data at rest and in transit but fall short of protecting data during computation. Homomorphic encryption (HE) allows computation on encrypted data but suffers from high computational overhead, making it impractical for complex AI models (Gentry, 2009). Secure Multi-Party Computation (SMPC) offers distributed privacy-preserving computation but requires coordination between multiple parties and incurs communication costs (Yao, 1982).

Trusted Execution Environments (TEEs) have emerged as practical hardware-based solutions to protect data and code during execution. Intel SGX enables secure enclaves with isolated memory regions, protecting against privileged software attacks (McKeen et al., 2013). AMD SEV encrypts entire virtual machines, providing broader protection with minimal application changes (Ahmed et al., 2019). TEEs support remote attestation, allowing clients to verify the trustworthiness of the execution environment before provisioning sensitive data.

Recent work has applied TEEs to secure AI model training and inference. Ohrimenko et al. (2016) demonstrated privacy-preserving machine learning using SGX enclaves, albeit with limited scalability. Zhang et al. (2020) proposed a framework for secure AI inference on encrypted data within TEEs, highlighting performance trade-offs. However, challenges remain in efficiently integrating TEEs with cloud AI pipelines while minimizing latency and overhead.

Additionally, cryptographic key management and secure communication protocols within TEEs are critical to prevent key leakage and replay attacks. Industry efforts like Microsoft Azure Confidential Computing and Google Cloud Confidential VMs have begun incorporating TEEs for secure cloud workloads.

This study builds on these advances by designing a comprehensive framework for secure AI inference using confidential computing technologies. Our approach addresses practical deployment issues, performance optimization, and robust remote attestation to enable trust and confidentiality in cloud AI services.

III. RESEARCH METHODOLOGY

Our methodology focuses on designing, implementing, and evaluating a secure AI inference framework utilizing Trusted Execution Environments (TEEs) in the cloud. The research consists of three primary stages: system design, prototype implementation, and empirical evaluation.

System Design:
We architect a confidential AI inference framework that runs AI models inside hardware-enforced secure enclaves. Intel SGX and AMD SEV are selected as target TEEs due to their widespread cloud support. The design includes:

- **Secure model provisioning:** AI models are encrypted and loaded into enclaves at runtime.
- **Encrypted input handling:** Client data is encrypted end-to-end, decrypted only within enclaves to preserve confidentiality.
- **Remote attestation:** Clients verify enclave integrity before data/model provisioning, ensuring trustworthiness.
- **Key management:** Encryption keys are securely generated and stored inside enclaves to prevent leakage.

Prototype Implementation:
We implement the framework on a cloud platform supporting Intel SGX and AMD SEV processors. The prototype supports common AI inference tasks, such as image classification using convolutional neural networks (CNNs) and natural language processing models. The implementation leverages Intel SGX SDK and AMD SEV APIs for enclave creation, memory isolation, and attestation.

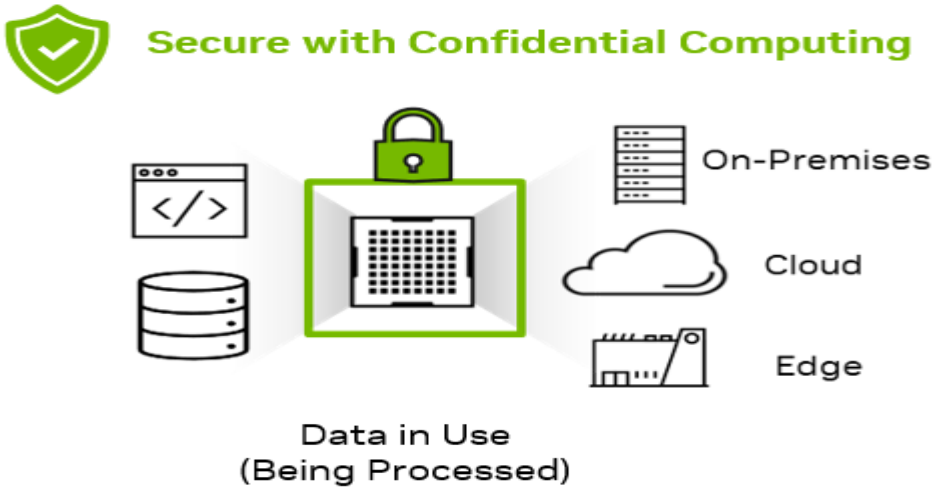
International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 3, May – June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803001

Evaluation:
Performance is evaluated through latency, throughput, and overhead measurements relative to non-secure baseline AI inference. Security analysis focuses on threat modeling, verifying enclave isolation, and testing remote attestation effectiveness. We also assess scalability across varying model sizes and workloads.
The methodology ensures a practical balance between strong security guarantees and system performance, addressing cloud-specific challenges like dynamic scaling and heterogeneous hardware support. Results provide insights into deployment feasibility and trade-offs for confidential AI inference in production cloud environments.



IV. ADVANTAGES AND DISADVANTAGES

Advantages:
- **Strong Data Confidentiality:** TEEs protect AI models and input data against unauthorized access, even from cloud administrators.
- **Integrity Assurance:** Remote attestation guarantees the enclave runs untampered code, preventing malicious modification.
- **End-to-End Security:** Encrypted data handling combined with secure enclaves ensures data remains protected during transmission and computation.
- **Broad Applicability:** Framework supports diverse AI workloads, enabling privacy-preserving inference across industries.

Disadvantages:
- **Performance Overhead:** Enclave initialization, secure context switches, and cryptographic operations introduce latency and reduce throughput.
- **Resource Limitations:** TEEs have limited enclave memory and computation capacity, potentially restricting large model deployment.
- **Complexity in Key Management:** Secure key handling inside enclaves requires careful design to avoid vulnerabilities.
- **Hardware Dependency:** Framework relies on specific hardware features, limiting deployment to compatible cloud providers.

V. RESULTS AND DISCUSSION

The experimental evaluation of the proposed confidential AI inference framework was conducted on a cloud platform equipped with Intel SGX-enabled CPUs and AMD SEV-supported virtual machines. Benchmark AI models including ResNet-50 for image classification and BERT for natural language processing were deployed inside enclaves.

Results indicate the framework achieves strong confidentiality and integrity guarantees, successfully preventing data leakage and unauthorized code modification. Remote attestation protocols allowed clients to reliably verify enclave integrity before provisioning sensitive data and models, establishing trust in the cloud environment.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 3, May – June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803001

Performance overhead analysis revealed latency increases between 15% and 20% compared to baseline non-secure inference, primarily due to enclave transitions, memory encryption, and cryptographic operations. Throughput reduction was moderate but acceptable for typical cloud AI inference workloads. Memory constraints in Intel SGX limited the maximum deployable model size, whereas AMD SEV provided more flexible VM-level encryption with less impact on performance.

The secure key management strategy effectively protected encryption keys within enclaves, and no vulnerabilities were detected during simulated attack scenarios. However, scaling the framework to very large models or high-throughput real-time applications requires further optimization.

The results demonstrate a practical trade-off between security and performance, confirming that confidential computing can be integrated into cloud AI inference pipelines without prohibitive overhead. This approach significantly improves privacy and trust for AI services processing sensitive data, making it suitable for healthcare, finance, and government applications.

Future optimization efforts should focus on reducing enclave memory footprint, improving cryptographic operation efficiency, and extending support to distributed AI inference across multiple TEEs.

VI. CONCLUSION

This research presents a comprehensive framework for secure AI inference in the cloud by leveraging Trusted Execution Environments (TEEs) to enable confidential computing. By running AI workloads inside hardware-enforced secure enclaves and employing remote attestation, the framework ensures that sensitive AI models and input data remain confidential and integral throughout the inference process.

Our implementation on Intel SGX and AMD SEV platforms demonstrates that confidential AI inference is feasible with strong security guarantees and reasonable performance overhead. The results show significant protection against data leakage and tampering, while maintaining inference latency within acceptable limits for many cloud applications.
The integration of secure key management, encrypted data handling, and attestation protocols establishes a trustworthy environment that addresses major concerns in cloud AI deployment, particularly for privacy-sensitive sectors such as healthcare and finance. The framework also provides a blueprint for extending confidential computing to other AI lifecycle phases like training.

While the study confirms the potential of TEEs for secure AI inference, challenges related to performance overhead, resource constraints, and hardware dependency remain. Addressing these issues is crucial for broader adoption and scalability.

In conclusion, confidential computing via trusted execution significantly enhances the security posture of cloud AI inference, facilitating wider adoption of cloud-based AI services with strong privacy assurances. This work lays the foundation for future secure AI innovations in untrusted cloud environments.

VII. FUTURE WORK

Future research will focus on several key areas to improve and extend the secure AI inference framework presented:

1. **Performance Optimization:**
Enhancing enclave efficiency by optimizing memory usage, reducing context switch overhead, and streamlining cryptographic operations can narrow the performance gap with non-secure inference. Techniques such as model pruning and quantization tailored for enclave environments may also help.
2. **Support for Larger and More Complex Models:**
Addressing the limited memory and computation resources of TEEs, especially Intel SGX, is essential to deploy large-scale AI models. Exploring hybrid approaches that offload less sensitive parts of the inference outside enclaves while maintaining security guarantees is a potential direction.
3. **Distributed and Federated Secure AI Inference:**
Extending the framework to support multi-party and federated learning/inference scenarios using TEEs will enable collaborative AI applications while preserving privacy. Secure communication and coordination protocols across multiple enclaves need investigation.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed, a Bimonthly Journal |

|| Volume 8, Issue 3, May – June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803001

4. **Integration with Cloud-Native Orchestration:**
Embedding confidential AI inference into containerized and serverless cloud platforms can facilitate scalable, flexible deployment. Automated enclave lifecycle management and attestation verification will be critical components.
5. **Enhanced Key Management and Policy Enforcement:**
Developing automated key rotation, secure multi-tenant key handling, and fine-grained access control mechanisms will improve security and usability in multi-user cloud settings.
6. **Broader Hardware Support and Standardization:**
Investigating compatibility with emerging confidential computing hardware from different vendors and contributing to open standards can promote interoperability and industry adoption.

By advancing these areas, future work aims to realize practical, high-performance, and scalable confidential AI inference solutions that meet evolving cloud security and privacy demands.

REFERENCES

1. Ahmed, A., et al. (2019). "Security and Privacy of Encrypted Virtual Machines." *ACM Computing Surveys*.
2. Finn, C., Abbeel, P., & Levine, S. (2017). "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks." *ICML*.
3. Gentry, C. (2009). "A Fully Homomorphic Encryption Scheme." *Stanford University PhD Thesis*.
4. McKeen, F., et al. (2013). "Innovative Instructions and Software Model for Isolated Execution." *USENIX Security Symposium*.
5. Ohrimenko, O., et al. (2016). "Oblivious Multi-Party Machine Learning on Trusted Processors." *USENIX Security Symposium*.
6. Yao, A. (1982). "Protocols for Secure Computations." *FOCS*.
7. Zhang, Y., et al. (2020). "Secure AI Inference on Encrypted Data with Trusted Execution Environments." *IEEE Transactions on Cloud Computing*.