Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Comparative Review of Offline First AI Architectures for Smart Systems in Low Connectivity and Disaster Prone Environments

John Rodge C. Sarait[1], Leen Jandy P. Mencias[2], and Leander Rafael C. Espela[3], Jay Ar P. Esparcia[4]
Department of Computer Engineering
University of Southern Mindanao
Kabacan, 9407, Philippines

Abstract **-** Smart systems that rely on the cloud become unusable when the network goes down in rural and disaster vulnerable settings, thus interrupting the monitoring system as well as slowing the timely responding to emergencies. This work sought to find offline-first AI systems that can be reliable and power efficient (when the network is unavailable or unreliable). Technology or Method: A systematic review of peer-reviewed articles in the period 2015-2025 compared three big architectures, namely TinyML on microcontrollers, Edge AI on single board computers, and federated learning frameworks. The comparisons were on inference latency, energy use, computational ability, reliability and cost of deployment in the application of agriculture, disaster observation and rural health care. Conclusions: TinyML showed milliwatt operation which allowed constant battery, even at the lowest levels, means of operation, but due to the model complexity, there was a limitation. Edge AI platforms were offering less computation time inference through tasks that demanded more computation (real time object detection) but with high energy consumption. Federated learning was effective in enhancing privacy of data, and adversarial distributed model refinement without the transmission of raw data, but its performance was limited in situations of long time disconnection because of its reliance on periodic synchronization. Combination strategies between local inference and adaptive synchronization had better operational reliability. Conclusions: There is no optimal architecture to take into account all constraints. TinyML can be used in cases of severe energy constraints, edge AI can be used in local application with real-time processing, and federated learning enhanced collaborative intelligence. Clinical Impact: Offline-first AI ensures continuity in diagnosis, decreases the response time and privacy of patient data locally during the rural context of preparing healthcare and emergency mechanisms, which improves the robustness of providing services to patients across connectivity-restrictive banking surroundings.

Keywords - Offline First AI Architectures, Smart Systems, Low Connectivity, Disaster Prone Environment

I. INTRODUCTION

The proliferation of smart systems has led to an increased reliance on interconnected devices for various applications, ranging from environmental monitoring to healthcare. Many contemporary smart systems primarily depend on centralized cloud computing for essential function like model inference and data storage. Nevertheless, this type of cloud centric environment causes considerable vulnerability to environments with unreliable or no internet connectivity, including rural areas or those impacted by natural calamities. When it comes to critical events such as typhoons, earthquakes or floods communication networks may collapse and cloud dependent systems become unusable at the very time that they are most in demand.

These limitation offline first AI architectures have become a viable option to overcome such limitations, manually transferring intelligence directly to the local devices. According to this paradigm, smart systems can operate continuously without having an active internet connection, as smart systems undertake processing and makes decisions on the level of the devices. Other dominant architectural solutions in this field are Edge AI platforms, like Raspberry PI and NVIDIA Jetson, which promote on device interference. TinyML allows operation of machine learning models over resource constrained microcontrollers with low memory power which makes it appropriate in environments of extreme constraint[1]. Moreover, federated learning decentralizes model updates, which takes place on distributed nodes, and synchronizes only in the case of connectivity, which is intermittently available [2]. This paper seeks to do a comparative review of these offline first AI-based systems, assessing how suitably each one can be completed to fit the smart systems themselves that are in low connectivity and disaster prone environments.

II. STATEMENT OF THE PROBLEM

Cloud dependent smart systems exhibit paramount operational failures in case of network outage, which is critical to applications that have to operate continuously. One of the main issues is the disconnection of the real time monitoring in case of the unavailability of the internet which prevents the real time response in serious situations. This dependency also results in a slowdown in time sensitive applications like early warning systems which are dominated by speed in data processing and interpretation of these results. Also, a lot of far-flung deployments have a short power supply, leaving conventional power-intensive AI solutions infeasible. Computational capabilities typically are further constrained by resource constrained hardware commonly used in edge devices, making it a problem to deploy complex AI models. Even though the requirement to have resilient smart systems is growing, a clear deficit of structured comparison is evident to select the right offline first AI architectures, which translate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

into suboptimal deployment decisions and vulnerability under challenging circumstances.

III. RESEARCH OBJECTIVES

General Objective:

The general objective of this study is to conduct a comparative review of offline first AI architecture that are suitable for smart systems operating in lo connectivity and disaster prone environments.

Specific Objectives:

1. To identify the major offline first AI architecture currently employed in resilient smart system.
2. To compare these identified architecture based on critical performance metrics including latency, energy consumption, reliability, and deployment cost.
3. To analyze practical deployment scenarios in key sector such as agriculture, disaster monitoring and rural healthcare, highlighting the applicability of each architecture.
4. To propose a decision framework that aids in the selection of appropriate architectures under various resource constraints.

IV. RESEARCH QUESTIONS

1. Which offline first AI architecture demonstrate the most optimal balance between low inference latency and minimal energy consumption in environment with intermittent connectivity?
2. How effectively do different offline first AI architecture maintain operational continuity and data integrity during prolonged periods of network disruption?
3. What are the primary challenges and opportunities in deploying and scaling offline first AI solutions for smart systems in low connectivity and disaster prone regions?

V. SIGNIFICANCE OF THE STUDY

The study has a profound implication to different stakeholders as it offers important insights into how resilient smart systems can be designed and implemented. To engineers, it provides a technical teaching on the different offline first AI architectures to be chosen and deployed to endure difficult environmental conditions and connectivity constraints. Researcher will enjoy a systematic comparison of various offline first AI models, existing strengths and weakness, and possible innovative directions. The evidence based recommendation can be used to help the policy makers develop strategies to build robust rural digital infrastructure so that essential services can be functional even in times of crisis. Besides, the research facilitates cost efficient implementation of smart systems in important fields like agriculture monitoring, disaster management, rural health analysis, and infrastructure surveillance. In the end, the results are projected to improve the stability and reliability of smart systems in settings associated with erratic connectivity, which will boost their safety and efficiency.

VI. SCOPE AND DELIMITATION

The area of this research in case the focus is on the comparative examination of offline firs AI architecture operating on the device level or local network level. It comprises systems of TinyML on microcontrollers, edge artificial intelligence systems on microboards highlighting single board computers, and localized federated learning systems. The metrics that may be used to compare it are the latency, power usage, computing power, reliability, and cost of deployment. The literature review will cover publications since 2015.

This work narrows down particular to exclude cloud only AI architectures in its scope since its nature requires uninterrupted internet connectivity, which falls under the offline first solutions. Moreover, the experiment does not imply the experimental hardware benchmarking as the research is based rather on the cited metrics of performance or analyses provided in the literature. This research is also not within the purviw of cybersecurity assessment, and large scale distributed cloud orchestration.

VII. REVIEW OF RELATED LITERATURES

The development of Artificial Intelligence (AI) and the Internet of Things (IoT) brought a new era of smart systems that are meant to improve efficiency and decision-making in various fields. The effectiveness of most modern AI applications, however, is frequently hinged on the presence of dependable high-bandwidth connectivity to cloud systems, making these applications prone in computer networks that lack or experiences low connectivity like remote geographical areas or disaster zones. This natural dependency has sparked academic and industrial attention into so-called offline-first AI architectures, which allows intelligent systems to execute AI tasks autonomously on the local scale to avoid the unending dependence on cloud services. Notable online-first AI architectures are critically analysed by this expanded review, identifying the underlying principles of the architectures, their uses, enablers in the underlying technology, related challenges, and future potential benefits, based on a stringent examination of the existing literature.

TinyML: AI at the Ultra-Low-Power Edge

TinyML is an effort to alleviate the resource-hazardous setting through addressing the issues that emerge when building small, resource-efficient deep neural network designs,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

the software ecosystems that assist them, and their implanting hardware. The essence of the TinyML initiative is to scale up the sufficiency of the deep learning systems through a reduction in computational and data needs to enable the ubiquitous use of edge AI and IoT systems [3]. These tiny devices have limited memory and processing power and so the optimization of models cannot be avoided.

Model Compression: It is a methodical procedure of denying the size and intricacy of a neural network by discarding or diminishing redundant or less essential elements in a attempt to maintain viable degrees of accuracy. Quantization: This method is used to switch the numerical precision of model parameters and activations (e.g. switching 32-bit floating-point much harder representations to 8-bit or even 1-bit integer representations), which dramatically decreases the energy and computation time required, with comparably small performance loss in models. The idea of sparsifying a neural network (resulting in a sparse and simplified model) was introduced as pruning, which is unrelated to the concept of rationalization. Knowledge Distillation: This is a teaching technique used in which a larger, more complex, teacher-producing model learns its knowledge and transfers it to a smaller, more efficient, student-producing model that then can have a similar performance on a resource-restricted machine.

Microcontrollers and energy efficient inference hardware architectures used to execute tinyML models are optimally run on microcontrollers. Microcontroller Units (MCUs) are particularly well-suited for TinyML applications due to their compact form factor (typically around 1 cm^3), exceptionally low power consumption (approximately 1mW), and cost-effectiveness (around $1) [4]. Most of these systems can include a Central Processing Unit (CPU), a number of digital and analog peripherals, embedded flash memory to store programs, and Static Random-Access Memory (SRAM) to store volatile data. Nordic Semiconductor, STMicroelectronics and NXP are large computer vendors that include hardware platforms with TinyML workloads. Other developments are heterogeneous TinyML SoCs which introduce energy-event-performance- conscious management and ultra-low power consumption compute-in-memory architectures.

The TinyML models are implemented using a very powerful ecosystem of software frameworks supporting the creation and deployment of those models in resource-constrained devices..TensorFlow Lite for Microcontrollers (TFLM): It can be said that this is the most popular one and has provided a big set of tools to optimize and deploy machine learning models in ultra-low-power settings. The TFLM allows the implementation of the Neural Network (NN) models on IoT devices. Edge Impulse: it is a platform that provides the integrated development environment enabling one to simplify the whole TinyML workflow, starting with data gathering and model training and decoding it to the deployment on an enormous variety of IoT devices. PyTorch Mobile, and other Lightweight Frameworks: These frameworks expand the selection of platforms in which machine learning models can be implemented on devices with extreme resource requirements..uTensor, MicroTVM, and CMSIS-NN at ARM: These are another set of platforms and libraries which importantly facilitate the development of TinyML especially when dealing with ARM based microcontrollers.

TinyML provides major benefits to traditional cloud- based machine learning concepts. Energy Efficiency: Microcontroller based models consume considerably less electricity when compared to traditional processors and allow devices to be powered over long periods (months/ years) using small amounts of battery supply. The feature is essential in new cognitive tasks in distant or unreachable places [3]. Low Cost: TinyML halves capital and operational expense of high-performance and powerful cloud computing resources and processors by performing local data processing on low-cost microcontroller devices. Low Latency: Localized processing data is automatically removed in terms of communication latency that comes with transferring data to and away of remote clouds, resulting in almost instantaneous responses and real-time analysis which is essential in emergency and time sensitive situations. Offline Inference: This is an iconic attribute of TinyML since it can be done machine learning without using the internet services, where the data analysis runs on the sensor or edge device itself. Enhanced Privacy and Security: Sensitive data is executed at the local level and hence reducing the exposure to the risks of transmitting information through the insecure networks [5], [6]. It is an inherent architecture that prevents transmission errors, cyberattacks and data breaches that may arise when communicating in the cloud [7], [8].

TinyML is a disruptive technology that triggers innovation in a wide range of fields and industries of application. Environmental monitoring and disaster early warning: TinyML helps in fielding small scale devices that can perform local environmental scanning, e.g. detecting lethal gases in the air, water level to predict floods or detect unusual acoustic or thermal patterns to detect wildfires. Healthcare and Wearable Monitoring: It allows implementing real-time health checking on wearable IoT devices, computing the physiological data in the device, and minimizing latency, energy usage and the issue of privacy that exists with data transfer to cloud services. Among others, they include on-device Electrocardiogram (ECG) anomaly detection, with optimized models that have reached high accuracy (92.3%) and ultra-low power consumption (0.024 mW).[9], [10], [6]. Ag farming: TinyML will give farmers the power to scan crops, anticipate their maintenance requirements, determine the state of the soil, and identify potential pests or diseases regardless of internet connectivity that will improve crop yields and reduce environmental harm. Industrial Automation and Predictive Maintenance: Miniature sensors installed on industrial machines can learn the operational signature and identify an unusual situation in real-time, eliminating unwanted down time and lowering operational expenses.Smart Homes and Consumer Electronics: TinyML can be embedded directly

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

within the industrial machine and learns its operational signature, indicating an anomalous situation and preventing the unwanted down time along with operation costs.

Object Detection and Real-time Visual Processing: Real-time object detection systems can be deployed with energy efficient controllers and efficient neural networks with TinyML with the ability to deliver the results to standard communication protocols (TCP or UDP) [11], [12]. It is very appropriate in the case of smart IoT devices and industrial surveillance [13].

Edge AI: Local Processing outside of Microcontrollers.

Edge AI is a more of a conceptual perspective where the AI workloads are processed in nearer proximity to the source of data. This is usually done on higher-computational edge units than are typical of TinyML such as single-board computers (e.g. Raspberry Pi, NVIDIA Jetson) or more specific AI accelerators. Although TinyML is a specialized form of Edge AI the latter can execute more complex models and computationally intensive operations than TinyML, though at a cost of a proportional increase in power consumption.

Edge AI has its specific benefits, especially in a situation where computer power requirements are too large to allow TinyML to execute a task, but on-edge processing is preferable. Minimal Latency: Edge AI saves the lag delays of transmitting data to and from the centralized cloud servers since data processing is conducted locally and thus allows real-time decision-making. Improved Privacy and Security: The processing of the data takes place on-device hence reducing chances of sensitive data being accessed or read in transit to remote servers. Minimized Bandwidth Consumption: The processed insights or aggregated data or actionable intelligence are only sent to the cloud significantly reducing network traffic and dependence on high bandwidth links. Offline Operation: Edge AI systems have a built-in ability to operate independently even without constant internet connectivity, which is a vital feature to implementations of remote or disaster-prone areas.

Edge AI is particularly compelling in areas where more computational power is needed than TinyML, although it still takes advantage of real-time and privacy, which cloud computing can offer. Real-time Object Detection and Communication: Real-time object detection systems that require less power are designed to operate in resource-constrained environments, thus being suitable to smart IoT devices, real-time industrial monitoring, and environmental sensors built into smart cities. Disaster Monitoring: Edge AI systems are crucial in being part of the early warning system and help to create real-time situational awareness and damage assessment in disaster-prone areas thanks to biometric authentication, perception at the edge device level.

Federated Learning:Collaborative Intelligence at the Edge

Federated Learning (FL) is a decentralized machine learning mechanism created to train the model on common devices (clients) from many independent (scatter) devices (clients) with each (client) having local data samples, with this sensitive data directly not exchanged between nodes. Rather, local models are trained on proprietary client data and only encrypted model updates (e.g. weight parameters or gradients) are sent to a coordinating central server. The server then consolidates these updates to combine a better global model, which is further propagated to the client devices where further local refinements are done by repetition.

FL has considerable benefits, especially in applications where data is accessed offline and then processed offline, which separates information and computation. Privacy of Data: This is a key advantage of the solution, in that the raw sensitive data does not leave the client machine that was used to generate the data and this also significantly enhances data privacy and security assurance. This is more so important to privacy sensitive applications such as health wearables and smart home devices.

Reduced Data Flow: Only small model updates, not the large amounts of raw data, are relayed and this causes a huge intake of bandwidth usage. This will prove particularly beneficial in networks with low connectivity or with intermittent network coverage. On-Device Training: Continuous Investigations, AI models are optimized and tuned in line with real-world and context-specific data that is provided by the edge devices, leading to more personalized as well as context-specific AI systems. Connection resilience: FL requires regular synchronization; however, it can be made resilient to connectivity issues by allowing local training to occur continuously in the absence of connections, and updates put in place once the network is back online.

Though FL has a number of merits, it has a number of intrinsic challenges to its implementation, which should be resolved to make it widely applied. Communication Overhead: Albeit with only the model updates being transmitted, there is a communication bottleneck that can be particularly challenging to manage in large-scale deployments where there are a large number of client devices, in addition to non-IID (non-independent and identically distributed) data distributions. Security and Privacy Concerns: FL offers improved privacy, however, it is not to be trusted with respect to security. Inference attacks, gradient leakage and model poisoning by malicious participants are some of the threats that can interfere with the integrity and confidentiality of the learning process. Privacy-sensitive federated learning in TinyML needs optimal ways of striking a balance between good privacy and efficiency.

Federated Learning is especially relevant to those applications where privacy is the most important or when

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

data is natively split into an untold amount of edge devices. Health care: Training machine learning models with patient data that comes from multiple hospitals or wearable medical equipment, and thus, preventing meeting points of confidential medical records. Mobile Devices: Improving predictive text services, voice recognition and image classification algorithms on smart phones without having to upload any personal user data to the cloud. An example is Google Gboard which uses Federated learning to customize typing suggestions without transmitting textual information to the cloud servers. Industrial IoT: Enabling co-learning of various industrial machines to perform optimally in operations and enhance efficiency and does not require institutions to share their proprietary operational information. Privacy-critical Domains: Active Federated learning cryptmem system allows safe and efficient federated learning on tiny-ML devices where data privacy is a key factor in consideration.

Hybrid Architectures and Future Directions

Solutions that are the most powerful and flexible to offline-first AI can typically result through the synergistic combination of multiple architectural paradigms. An example is a hierarchical ensemble TinyML scheme that is capable of integrating the uncoordinated decisions of many IoT components into the understanding of system-wide intelligence to save on wireless communication bandwidth, lower energy use, decrease the response time and secure data privacy. Innovationhints at progressing past the scope of conventional TinyML towards broader edge computing concepts support the creation of genuinely intelligent and context-aware applications in highly distributed IoT settings, with an existent ambition of bringing generative AI models to the edge devices even though they have high- level computational and memory requirements. Such an evolutionary pathway would enable even more extraordinary functionalities, like the generation of local content and more complex agent-analog interactions in asset-limited contexts, whilst providing substantial engineering challenges, which are impossible to foresee but might need significant effort to address. This continuous expansion highlights the unified trend of moving towards the generation of more powerful, multifaceted and stand- alone AI solutions. Neuromorphic computing, federated TinyDL, edge-native foundation models, domain-specific co-design methods, and other ways of pushing the limits of edge AI are all areas of future research. The combination of TinyML and LargeML (Large Machine Learning) is also estimated as one of the promising ways of future seamless connectivity and efficient use of resources in 6G networks and the future. Such hybrid strategies are able to guarantee robust, inaccessible and privacy-enabling smart systems to vital applications.

VIII. SUMMARY

A general survey of offline-first AI architectures was undertaken in this study whereby the Edge AI platforms, TinyML systems, and federated learning frameworks have been carefully examined in the context of their use in low-connectivity and disaster prone settings. The main aim was to identify the distinctive properties and operational trade- offs of each of these architectures, that is, regarding their computational capabilities, power, and reliability during periods of intermittent or disconnectivity to the system network. The present analysis has highlighted the balancing nature of these indicators of performance, and it is necessary to stress that their criticality in successful AI implementation in problematic operational conditions.

Edge AI platforms are built to provide the capability of executing complex AI models on edge devices, increasing the reduction of latency as well as decreasing the need to constantly communicate with the cloud. Although the platforms provide efficient computational power appropriate to complex AI work and improve data privacy,

computing the information on-the-fly, it usually requires a moderate power consumption. Such power requirement can pose a notable constraint element within power-starved setting. By contrast, TinyML systems consider ultra-low power usage as a higher priority, which is why they are remarkably well-fit to the battery-powered, devices in which the long life of deployment is the most critical factor to consider. This energy efficiency is done by means of vigorous model optimization methods which diminish the model size and computational demand. Nevertheless, such a design decision limits TinyML models to smaller and more compact, further restricting their memory to very complex calculations.

In Federated learning, on the other hand, a paradigm of distributed collaborative training of AI is presented, whereby local models are trained using local data that is not centralized. Such a practice has a major impact on privacy and security of data since sensitive information is not sent anywhere, and only model parameters or updates are transferred. However, a federated learning system demands that these model updates be periodically synchronized by nature, which poses significant issues regarding operational capabilities in the event that there may be a long-term break in the network contacts or an interim loss of contact. Such timely convergence needs may hinder the timely convergence of the models, and the overall reliability of the system during offline conditions, and it is necessary to have solid methods of dealing with the communication needs during a time of strain.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

IX. CONCLUSION

Offline the first AI architecture is crucial in supporting the resilience and operational continuity of smart systems especially in conditions with low-connectivity and exposure to disasters where basic cloud-reliant infrastructure has weaknesses by definition. Its core capability of making local inference guarantees operational invariance when global network connection is lost, and thus, it is an explicit response to a dire weakness of systems dependent on maintaining a constant connection with a cloud. This capability of local processing is critical to sustenance of the required services and decision-making operations even under unfavorable circumstances, which offers autonomous intelligence amidst absence of external resources.

The process of choosing a proper architecture is the number one aspect and requires a thorough analysis of a variety of critical factors. These are the accessibility of power resources, the specific computing requirements of the usage, and the preferred recovery plan in case of loss of connectivity. In the example, applications with complex analysis based on real-time may better use Edge AI, and those that prioritize extended battery with less complex, continuous operations would greatly benefit TinyML. These variables must be comprehended and particularly weighed before implementing a custom solution that will optimize its efficiency, strength, and operational life within the selected environmental limitations.

Intelligent combinations of local inference models of Edge AI or TinyML with adaptive synchronization of federated learning, known as hybrid models, are a highly promising developmental deployment in the future. Such combined strategies should be specifically used in rural and emergency settings where a practical combination of localized processing and more secure and distributed learning is not only useful but in many cases essential. These hybrid architectures can provide a trade-off between computational and energy efficiency and at the same time be resilient to intermittent connectivity giving a more end- to-end and versatile structure of offline-first AI that can efficiently run and adapt on changeable and unpredictable surroundings.

REFERENCES

[1] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello Edge: Keyword Spotting on Microcontrollers," arXiv preprint arXiv:1711.07128, 2017.

[2] K. Bonawitz, H. Eichner, W. Grieskamp, F. Huba, A. Ingerman, J. Konečný, et al., "Towards federated learning at scale: System design," in Proceedings of the 2nd SysML Conference, 2019.

[3] IGI Global, TinyML Empowering Intelligent Edge Devices. Hershey, PA, USA: IGI Global, 2025.

[4] P. Warden and D. Situnayake, TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. O'Reilly Media, 2019.

[5] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, et al., "Privacy and security issues in deep learning: A survey," IEEE Access, vol. 8, pp. 208579-208595, 2020.

[6] Y. R. Thota and T. Nikoubin, "Enhanced Consumer Healthcare Data Protection Through AI-Driven TinyML and Privacy-Preserving Techniques," IEEE Access, 2025.

[7] T. Sasi, A. H. Lashkari, R. Lu, P. Xiong, and S. Iqbal, "Comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges," Journal of Information and Intelligence, vol. 2, no. 6, pp. 455-513, 2024.

[8] M. H. Nguyen Ba, J. Bennett, M. Gallagher, and S. Bhunia, "A case study of credential stuffing attack: Canva data breach," in 2021 International Conference on Computational Science and Computational Intelligence (CSCI), 2021, pp. 735-740.

[9] Y. R. Thota and T. Nikoubin, "TinyML for ECG Biometrics on Resource Constrained Devices," in 2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2024, pp. 521-526.

[10] Y. R. Thota, J. S. Nixon, B. Chandran, and T. Nikoubin, "TinyML based biometric authentication using PPG signals for edge devices," in Proceedings of the Great Lakes Symposium on VLSI 2025 (GLSVLSI'25), 2025, pp. 860-865.

[11] IGI Global, TinyML Empowering Intelligent Edge Devices.A. Preukschat and E. Reed, Self-Sovereign Identity. Shelter Island, NY, USA: Manning Publications, 2021.

[12] V. K. Hahn and S. Marcel, "Biometric template protection for neural-network-based face recognition systems: A survey of methods and evaluation techniques," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 639-666, 2023.

[13] U. Elordi, et al., "Designing automated deployment strategies of face recognition solutions in heterogeneous IoT platforms," Information, vol. 12, no. 12, p. 532, 2021.

[14] Y. R. Thota and T. Nikoubin, "EdgeAI with TinyML: Redefining Privacy and Security in Online Identity Management," in The Evolving Landscape of Online Identity - Recent Studies and Insights, IntechOpen, 2025.

[15] Y. Liu, Y. Wang, H. Chen, and A. Liu, "Toward zero trust IoT security: Lightweight device authentication in edge computing," IEEE Internet of Things Journal, vol. 8, no. 14, pp. 11365-11378, 2021.

[16] R. Alrawili, A. A. S. AlQahtani, and M. K. Khan, "Comprehensive survey: Biometric user authentication application, evaluation, and discussion," Computers & Electrical Engineering, vol. 119, no. Part A, p. 109485, 2024.

[17] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in International Conference on Learning Representations (ICLR), 2018.