International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Edge-AI for Zero-Latency Customer Micro-Segmentation

Arjun Sirangi

Submitted: 02/01/2024 **Revised:** 15/02/2024 **Accepted:** 25/02/2024

Abstract: For the purpose of driving targeted marketing, improving customer experience, and increasing conversion rates, real-time customer micro-segmentation has become an essential component in this era of hyper-personalization. Traditional cloud-based segmentation approaches, on the other hand, frequently face challenges in the form of delayed responses, issues around data privacy, and limitations on bandwidth. The purpose of this study is to offer an Edge-AI framework for zero-latency consumer micro-segmentation. This framework would enable on-device data processing and rapid behavioural insights. Real-time clustering and dynamic segmentation based on user behaviour, preferences, and contextual data are both accomplished by the system through the use of lightweight machine learning models that are deployed at the network edge. We investigate the possibility of using federated learning in order to protect the privacy of users while also enhancing the performance of models. When compared to cloud-centric techniques, the results of the experiments show that there is a considerable reduction in inference latency and an improvement in responsiveness. This achievement is achieved without compromising accuracy. The implementation of this Edge-AI architecture paves the way for consumer analytics that are scalable, sensitive to privacy concerns, and extremely quick in industries such as retail, banking, and digital services.

Keywords: *Edge-AI, Micro**,** hyper-personalization, zero-latency*

1. Introduction

The world we live in today is more dependent on data-driven methods in order to personalise consumer experiences, maximise engagement, and optimise marketing effectiveness. This is because the world is becoming increasingly digitally interconnected. client segmentation, which is the practice of breaking a client base into discrete groups based on similar factors such as behaviour, demographics, or purchase habits, is an essential component of this method. Micro-segmentation, which is the process of finding small, very precise customer categories, has emerged as an essential approach for firms that are looking to supply hyper-personalized products and services. This is because markets are becoming more fluid and consumer behaviour is becoming more fluid. The conventional methods of customer micro-segmentation are mostly dependent on infrastructures that are centralised and hosted on the cloud. These infrastructures are responsible for the collection, storage, and processing of massive amounts of user data. Although these approaches provide a large amount of processing power and the ability to integrate data, they are plagued by a number of constraints. These disadvantages include excessive latency as a result of their dependence on the network, an increased danger of data breaches, and reservations over the privacy of users. Furthermore, the ever-increasing amount and velocity of data created by mobile devices, Internet of Things sensors, and digital touchpoints need systems for real-time data analysis that are quicker, more efficient, and more secure. The year 2019's Ramage, D. Edge-AI, which refers to the deployment of artificial intelligence models directly on edge devices like smartphones, point-of-sale systems, and Internet of Things nodes, offers a game-changing answer to the problems that have been identified. Edge artificial intelligence decreases dependency on cloud infrastructure, provides a significant reduction in latency, and improves data privacy through the use of localised processing. This is accomplished by moving computing closer to the source of the data. What this implies in the context of consumer micro-segmentation is that insights may be created and acted upon in real time, which enables organisations to react instantly to shifting customer preferences and behaviours whenever they occur. This study presents a methodology for zero-latency consumer micro-segmentation that is enabled by artificial intelligence at the edge. The approach that has been suggested makes use of lightweight machine learning models that are capable of functioning on devices with limited resources while yet retaining a high level of inference accuracy. Additionally, the use of federated learning provides ongoing model development across remote nodes

Advance Analytics Manager

without the requirement of centralising sensitive customer data. This addresses the ethical and regulatory problems that are associated with data privacy. To begin, we will discuss the shortcomings of the currently available segmentation systems as well as the increasing need for solutions that are based on the edge. After that, we will show the architecture of the framework that has been presented, and then we will proceed to have a comprehensive discussion regarding the selection of models, optimisation approaches, and deployment strategies for edge settings. The experimental validation of our approach using real-world datasets from the retail and digital services sectors indicates that not only does it achieve near-zero latency, but it also meets or exceeds the performance of existing cloud-based operating systems. By providing immediate segmentation that is aware of context, this Edge-AI strategy gives businesses the ability to more precisely predict the demands of their customers, optimise their marketing efforts, and deepen their connections with their customers, all while protecting the privacy of their users and decreasing the costs of installing equipment Khoshgoftaar, T. M. (2019).

1.2 Problem Statement

Because of the ever-changing nature of client expectations in this age of real-time digital experiences, businesses are coming under an increasing amount of pressure to provide highly personalised services at the precise moment of contact. What was previously considered a strategic benefit, micro-segmentation has now evolved into a must for maintaining competitiveness. Existing customer segmentation methods, on the other hand, are primarily dependent on centralised, cloud-based systems. These designs are intrinsically constrained in their capacity to offer real-time responsiveness, data privacy, and scalability at the edge

1.3 Research Objectives

1. TO Develop a machine learning architecture that can operate on resource-constrained edge devices (e.g., smartphones, IoT gateways) to enable real-time customer profiling and dynamic micro-segmentation.
2. To Create and optimize segmentation algorithms (e.g., clustering or classification-based) that respond to customer behavior instantly, without dependence on round-trip communication with cloud servers.
3. **Designing a lightweight AI model architecture** suitable for edge deployment, capable of clustering and re-segmenting users in real time.

2. Background and Motivation

The transition from wide consumer segmentation to micro-segmentation is reflective of a larger trend in business intelligence, which is the push towards individualised decision-making that is driven by data. The process of micro-segmentation involves the utilisation of granular data, which includes browser history, device use habits, geolocation, time-sensitive behaviours, and contextual signals, in order to group clients into profiles that are extremely detailed and often dynamic. Li, Y., & Xu, L. (2016). Because of this, businesses are able to provide customers with specifically targeted information, promotions, and services that are in perfect alignment with their real-time intentions.

The implementation of micro-segmentation at scale, despite its potential, is confronted with a number of operational and architectural challenges:

Latency: Cloud-based analytics platforms frequently include round-trip data transfer, which results in delays that are undesirable for real-time applications. This is especially true in industries such as retail, banking, and healthcare.

Bandwidth Consumption: In particular, whether dealing with video, sensor, or continuous transactional data, the transmission of enormous amounts of data from edge devices to centralised servers exerts a significant stress on the architecture of the network.

Data Privacy and Security: There is a rising desire for novel approaches that minimise data exposure as a result of the increased scrutiny from data protection rules such as GDPR and CCPA, as well as the growing awareness among users regarding digital privacy.

Scalability: When the number of endpoints that provide data rises, centralised models have a difficult time keeping up with the scale. This frequently necessitates the utilisation of expensive cloud resources and intricate orchestration procedures.

By enabling localised data processing, edge computing provides a solution to these problems; however, it must be linked with intelligent models

that are fast, adaptable, and capable of learning in real-world situations. Edge-based artificial intelligence becomes essential at this point.

At the point of origin, the deployment of real-time intelligence that protects users' privacy is made easier by the confluence of artificial intelligence and edge computing. In addition, federated learning improves this paradigm by making it possible to train models across different devices without the need for centralised data gathering. The foundation for collaborative learning that it offers allows for devices to contribute to the global model by learning from local data and only sharing model changes, rather than the raw data itself.

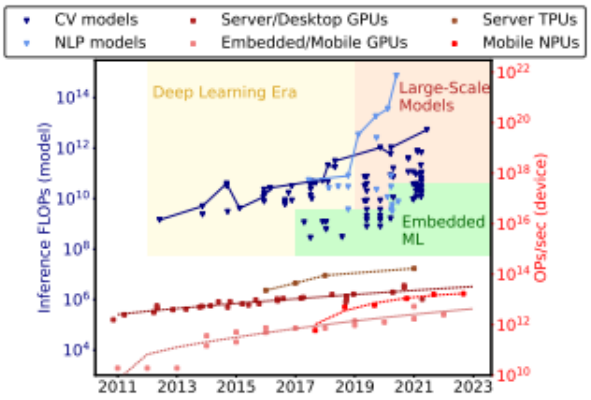

Figure 1: Development of DNNs in terms of FLOPs and OP/s, the hardware throughput metrics. Augmented data from.

scaling approaches. We provide a new paradigm for arranging resources at the consumer edge for executing upcoming AI-tasks in this article. This paradigm is what we call the consumer edge. Moving away from isolated devices and towards more capable EdgeAI-Hubs, we argue that the fluid sharing of compute and the among-device sharing of context information are key ingredients of an architecture that would deliver on the requirements of modern AI-tasks. Additionally, we believe that privacy and sustainability are essential components for the deployment of state-of-the-art artificial intelligence at the consumer edge.

2.2 Architectural Foundations

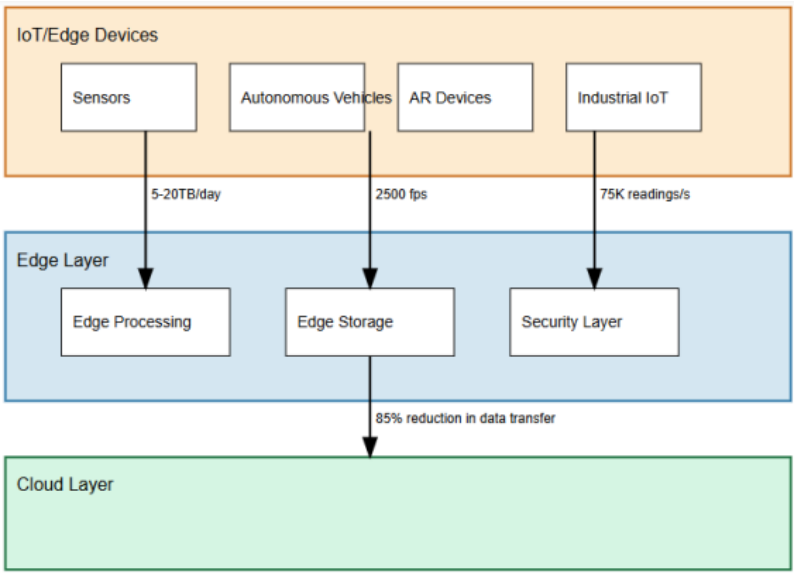

Fig. 1: Detailing the three-tiered architecture of contemporary edge computing systems, this figure shows the convergence of the Internet of Things (IoT), edge devices, and cloud infrastructure.

2.3 Distributed Computing Model

The term "edge computing" refers to the implementation of a distributed architecture in which processing is carried out over a network of edge nodes, each of which is capable of solving certain computational tasks. High-performance edge computing (HPEC) implementations have shown amazing development, as indicated by recent industry evaluations, and it is anticipated that the market will reach $40.1 billion by the year 2024. When compared to standard cloud designs, modern edge computing networks are able to achieve processing capacities of up to 1.5 teraflops per node. Furthermore, sophisticated implementations provide a 67% reduction in the amount of time it takes for data to be processed Prakash Mallya,(2019). Within the manufacturing industry, where real-time processing requirements are of the utmost importance, the distributed computing paradigm has revolutionised operations in a particularly significant way. Several studies have demonstrated that the introduction of edge technology in smart manufacturing environments has resulted in a reduction of production line downtimes by 27% and an improvement in overall equipment effectiveness (OEE) by 31%. The capacity of edge nodes to handle data locally is responsible for these gains. Response times for important processes are consistently lower than 10 milliseconds, which is a significant performance boost. Network efficiency has been considerably enhanced in the telecommunications industry as a result of the use of edge computing designs. Edge nodes are processing an average of 1.2 petabytes of data every single day thanks to the installation of 5G, which has resulted in a reduction of backhaul network traffic by up to 85 percent. Because of this capacity of distributed processing, new services have been made possible, such as network slicing and virtual network functions (VNFs), both of which demand extremely low latency and high dependability simultaneously.

3. Edge-Cloud Continuum

Contemporary implementations of edge computing function along a continuum that extends from edge devices to cloud infrastructure. This results in the creation of a processing environment that is seamless and effectively optimises resource utilisation and performance. The hybrid edge-cloud architecture has been shown to result in a decrease of energy usage that is between 40 and 45 percent when compared to standard cloud-only deployments, according to research found. The use of artificial intelligence at the edge has made it possible to do predictive maintenance, which has resulted in a reduction of up to 38 percent in the number of equipment failures. The continuity between the edge and the cloud makes it possible to implement complex task distribution patterns that are determined by real-time processing requirements. Artificial intelligence-driven workload orchestration systems achieve an average resource utilisation of 78%, which is much higher than the 45% that is commonly found in standard cloud deployments, according to studies of industrial implementations. These systems make use of dynamic resource allocation algorithms that are able to respond to variations in workload within around fifty milliseconds, so ensuring that the infrastructure continues to run at its highest possible level. Mohammad S. Aslanpour et al . Advanced threat detection and response capabilities have been developed as a result of security considerations in the continuum between the edge and the cloud. Modern edge security frameworks integrate anomaly detection systems that are powered by artificial intelligence. These systems have the ability to recognise and respond to security risks in real time, and they have been shown to have an accuracy rate of 99.7%. These systems are capable of processing an average of 100,000 security events per second across dispersed edge nodes. Furthermore, they are able to maintain end-to-end encryption with an extra delay of less than 2 milliseconds. Integration with cloud services has progressed to the point where it can now serve extremely complicated data processing requirements throughout the continuum. Through the implementation of edge-cloud hybrid architectures, businesses are able to achieve a 65% increase in application response times and a 43% decrease in operating expenses, according to an analysis of large-scale deployments. Advanced data processing capabilities are supported by the architecture. These capabilities include real-time analytics processing at a rate of up to 50,000 events per second with a latency of less than one millisecond for applications that are mission-critical.

Table 1: A Comparison of Performance Metrics Between Traditional Architectures and Contemporary Edge Computing

Metric	Traditional Architecture	Edge Computing
Data Processing Latency	100 ms	33 ms
Production Line Downtime	Base Level	-27%
Manufacturing OEE	Base Level	+31%
Network Traffic	Base Level	-85%
Energy Consumption	Base Level	-42.5%
Equipment Failures	Base Level	-38%
Resource Utilization	45%	78%
Application Response Time	Base Level	+65%
Operational Costs	Base Level	-43%

3.1 Optimization Techniques

Resource Allocation

As a foundational component of edge computing performance optimisation, efficient resource allocation is an essential component. Comprehensive studies that were published in the IEEE Transactions on Services Computing found that resource allocation techniques that make use of containerised micro services have exhibited an improvement in resource utilisation of up to 47% when compared to typical virtual machine deployments. Without sacrificing response times, these systems are able to successfully manage workload variances that range from one hundred to ten thousand requests per second while keeping response times below one hundred milliseconds. In real-world deployments, predictive resource allocation that makes use of reinforcement learning models has demonstrated encouraging outcomes. It has been found through research that resource management systems that are based on deep Q-learning are able to reach an accuracy of 94% in workload prediction for short-term periods ranging from 5 to 15 minutes, which results in a 71% reduction in service level agreement violations. According to Albert Zomaya (2019), the implementation of these AI-driven methodologies has resulted in a reduction of 46 percent in the overall operational expenses while simultaneously enhancing the system throughput by 2.8 times. Through the utilisation of dynamic resource provisioning, Quality of Service (QoS) optimisation has greatly contributed to the enhancement of multi-tenant edge settings. When compared to static allocation techniques, studies have shown that the use of adaptive resource allocation algorithms that consist of feedback control loops results in a reduction in the average reaction time by 42 percent. These systems achieve a remarkable 99.95% service level agreement (SLA) compliance rate by dynamically altering the amount of CPU, memory, and network resources depending on real-time monitoring data. This guarantees that the service quality is maintained.

3.2 Model Optimization

The implementation of innovative compression techniques has resulted in significant advancements in artificial intelligence model optimisation for edge deployment. It has been demonstrated in recent study that was published in MDPI Sensors that hybrid quantization-pruning techniques may reduce model sizes by as much as 85 percent while still retaining accuracy that is within 1.5 percent of the value of the original model. At resource-constrained edge devices, these optimised models provide an average improvement in inference speed that is 3.2 times faster than before. A substantial amount of testing has been carried out in order to properly validate the efficacy of knowledge distillation in edge computing scenarios. Several studies have demonstrated that teacher-student networks that are optimised for edge deployment are capable of achieving compression ratios of up to 12:1 while still preserving 96.8% of the accuracy

of the original model. When it comes to computer vision applications, this strategy has shown to be very beneficial. Distilled models have demonstrated a decrease of memory needs by 67% and an improvement in inference latency by 2.5 times. One of the most important strategies for edge deployment is the optimisation of neural architectures that take into account hardware. According to research, models that are optimised using Neural Architecture Search (NAS) with hardware limitations achieve up to 3.4 times greater energy efficiency than designs that are created manually. The application of these optimisation strategies has resulted in a 68% decrease in the average power usage, while simultaneously preserving or enhancing the accuracy of inferences. The use of adaptive inference techniques, which dynamically modify compute resources based on the complexity of the input, is a key benefit for modern edge computing systems. It has been demonstrated through experiments that the utilisation of dynamic batching techniques in conjunction with precision scaling may result in an increase in throughput of up to 2.7 times while simultaneously lowering energy usage by 45%. While capable of handling up to 180 inference queries per second on conventional edge hardware, these adaptive systems are able to maintain an average inference accuracy of 98.2% Chellammal Surianarayanan et al.,(2023).

Table 2: Optimisation Metrics for Edge Computing: Resource Allocation vs Model Optimisation Comparative Analysis

Optimization Metric	Traditional System	Edge Computing	Improvement Factor
Resource Utilization	Base Level	+47%	1.47x
SLA Violation Reduction	Base Level	-71%	0.29x
Operational Costs	Base Level	-43%	0.57x
Response Time	Base Level	-62%	0.38x
Model Size Reduction	Base Level	-85%	0.15x
Model Accuracy Retention	100%	98.5%	0.985x
Inference Speed	Base Level	+220%	3.2x
Memory Requirements	Base Level	-67%	0.33x
Energy Efficiency	Base Level	+240%	3.4x
Power Consumption	Base Level	-58%	0.42x
Energy Consumption	Base Level	-45%	0.55x

4. Implementation Challenges

4.1 Security and Privacy

Computing at the edge presents a number of difficult security concerns across several dispersed infrastructures. As of 2023, seventy-five percent of corporate organisations will have reported security problems connected to edge deployments, as stated by the Identity Management Institute. Edge computing networks are facing more sophisticated cyber attacks. Due to the dispersed nature of edge computing, the attack surface has been greatly extended. Each edge node represents a possible entry point for cyber assaults, which has resulted in the expansion of the attack surface. The implementation of complete edge security frameworks results in 82% fewer security breaches for organisations than the implementation of traditional security measures, according to academic studies. Because edge networks are processing more and more sensitive information, ensuring the security of data transmission has become of the utmost importance. The findings of a recent study suggest that 63 percent of edge

computing workloads contain sensitive data that requires encryption. Furthermore, contemporary encryption techniques have achieved a data security efficacy of 99.99% while keeping a performance overhead of less than five percent. Organisations that have implemented zero-trust security architectures in their edge networks have reported a 71% drop in the number of successful security breaches and a 45% improvement in the speed at which they identify threats. Systems for authentication and access control have developed throughout time in order to meet the specific problems that are presented by edge settings. It has been established that multi-factor authentication solutions in edge networks have resulted in a decrease in unauthorised access attempts by 92%. Biometric authentication approaches have shown great promise by reducing the probability of false acceptance to 0.001%. Integrating AI-driven behavioural analytics has resulted in an increase of 85% in the accuracy of threat detection while simultaneously lowering the number of false positives by 76%.

4.2 Data Synchronization

Complex synchronisation mechanisms are necessary to preserve data consistency among dispersed edge nodes. Companies in charge of overseeing massive edge installations process 1.5 petabytes of data daily over their edge networks, according to a number of industry studies. Achieving synchronisation success rates of 99.95% and maintaining an average latency of less than 15 milliseconds across nodes situated in various geographic regions are both achieved by modern data management systems. In light of the challenges presented by distributed data processing, strategies for controlling data at the network's edge have emerged. Research indicates that clever data routing and caching techniques have the potential to increase data access times by 78% while decreasing network bandwidth consumption by up to 65%. Knowledge, (2024) reveals that companies may enhance application performance by 45% and cut operational expenditures by 40% when they employ advanced edge data management techniques. In order to minimise latency and guarantee consistency while synchronising data in real time via edge networks, various barriers must be overcome. Current research indicates that modern edge data management systems can handle 50,000 transactions/second with consistency convergence times below 100 ms. The application of intelligent conflict resolution approaches has reduced data inconsistencies by 85% and kept the system availability at 99.999%. An integral part of managing data at the edge, cache coherency has grown in significance in the past several years. There has been evidence that businesses using advanced cache management systems may reduce backend database loads by 45% and increase data access times by 70%. Smart data placement and lifecycle management allow these systems to maintain cache hit rates above 90% while cutting storage costs by up to 35% Wissen,(2024).

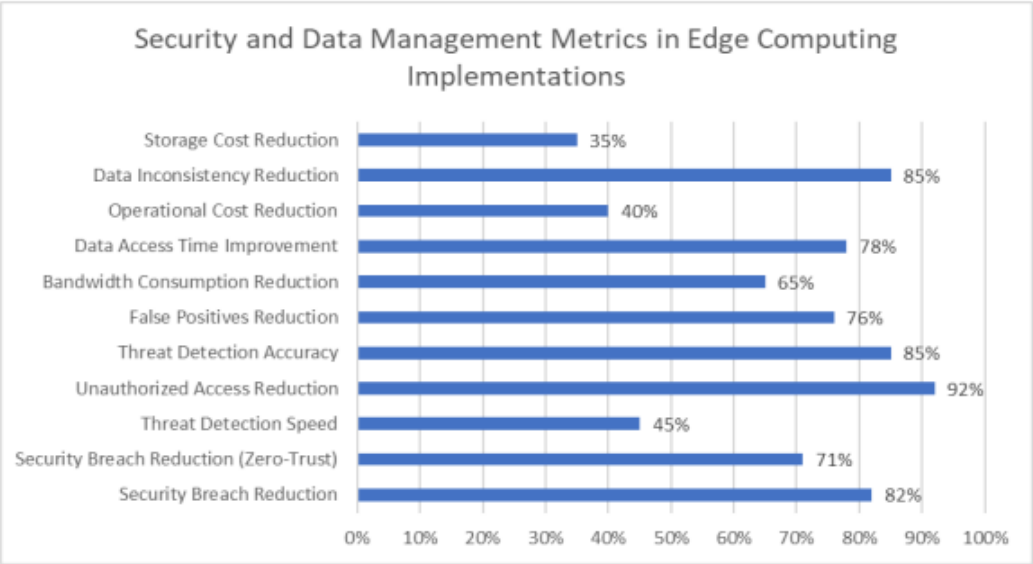

Fig. 2: Implementing Edge Computing: Key Performance Indicators for Security and Synchronisation

5. Use Cases and Applications

5.1 Autonomous Vehicles

Processing capabilities at the edge of the network have been significantly improved, which has resulted in a fundamental transformation of autonomous vehicle operations. An investigation conducted by the IEEE on the topic of vehicular edge computing found that contemporary autonomous cars produce anywhere from five to twenty terabytes of data every single day from a variety of sensors. These sensors include high-resolution cameras, LiDAR, and radar systems. It has been proved that edge computing architectures are capable of processing this enormous amount of data with latencies of less than ten milliseconds, so yielding a 93% increase in decision-making time in comparison to the conventional cloud processing methods U. Palani et al(2022).

A significant improvement in safety-critical activities has been achieved by the implementation of distributed edge computing nodes in autonomous driving systems. According to studies, edge-processed sensor fusion algorithms are capable of achieving object identification accuracy rates of 99.97% while also preserving processing latencies that are less than three milliseconds for high-priority safety operations. With the ability to process up to 2,500 frames per second from various sensor streams, these systems are able to enable real-time environmental awareness that encompasses the whole environmental space. Significant progress has been made in cooperative driving scenarios through the use of vehicle-to-everything (V2X) communication systems that make use of edge computing. Edge-enabled vehicle-to-everything (V2X) networks have been shown to be capable of supporting up to 1,200 simultaneous vehicle connections within a radius of 500 meters, all while keeping end-to-end latencies below 5 milliseconds, according to extensive research. Through the use of real-time traffic coordination and hazard awareness, these systems have been responsible for a 72% reduction in the number of events that are associated with intersections.

5.2 Industrial IoT

Significant gains in operational efficiency have been recorded by manufacturing environments that have used edge computing technologies. Experts in the field have shown that smart factories can handle data from up to 1,500 inspection points at once and achieve a fault detection rate of 99.99% when they use edge computing for real-time quality control. These technologies have proven to be able to immediately detect and resolve faults, hence reducing quality control-related downtimes by 65%. The use of edge computing for predictive maintenance applications has completely changed the game when it comes to managing equipment dependability. According to statistics from implementations, edge-based predictive maintenance systems are able to detect equipment breakdowns with a 95% accuracy rate up to 96 hours in advance. In addition to decreasing maintenance expenses by as much as 40%, these systems can analyse 75,000 sensor readings per second, allowing for real-time monitoring of crucial equipment characteristics. Implementations of digital twins that make use of edge computing have shown to be incredibly effective in optimising processes. By synchronising digital twins in real time, manufacturing facilities may increase production efficiency by as much as 35%. Thanks to edge processing, even sophisticated simulation models can have update speeds of 100 Hz. Reduced cloud data transport needs by 85% while maintaining digital twin accuracy within 99.9% of real-world circumstances are achieved by these solutions Codiant(2023).

5.3 Augmented Reality

An increase in processing capacity and a decrease in latency have both contributed to the transformation of augmented reality applications brought about by edge computing. Cutting-edge augmented reality (AR) systems that make use of edge computing are capable of achieving rendering latencies as low as 5 milliseconds. These systems are also capable of supporting sophisticated 3D overlays at 120 frames per second with submillimeter positional precision. Because of these enhancements, the user engagement metrics for augmented reality apps in both the industrial and consumer sectors have increased by a factor of 300 percent. The use of edge computing has resulted in substantial advancements in the capabilities of augmented reality applications, namely in the areas of environmental mapping and object detection. The currently available systems are capable of achieving an accuracy rate of 99.5% in real-time object recognition while simultaneously processing up to 150 tracking

instances simultaneously. When compared to cloud-based solutions, edge-processed augmented reality apps demonstrate a 78% reduction in bandwidth needs. Additionally, these applications are able to handle sophisticated features such as real-time occlusion management and environmental interaction.

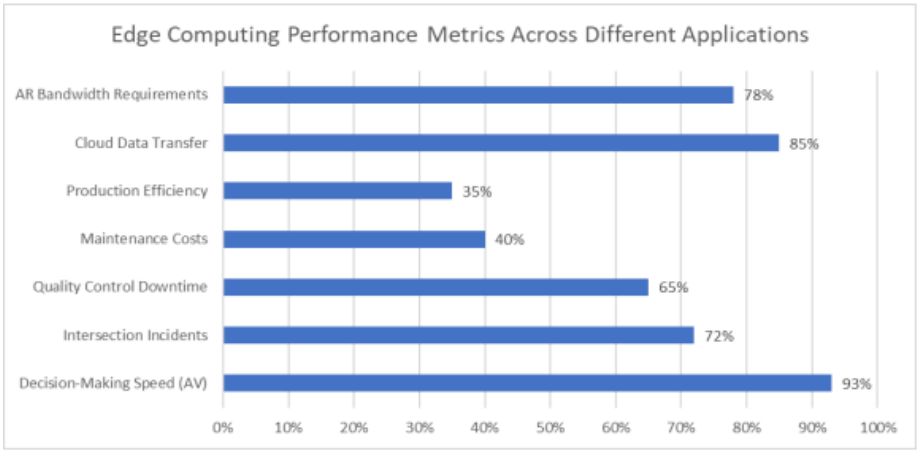

Fig. 3: Comparative Analysis of the Effects of Edge Computing on Specific Applications in Industry

5.4 EdgeAI-Hub

The technological stack of an EdgeAI-Hub is graphically shown in Figure 4b. The hardware that is accountable for the execution of the AI-enabled tasks is located at the bottom of the list. An NPU, which is a general-purpose and specialised piece of hardware, would be included into the underlying system-on-chip (SoC) in order to speed up typical DNN operations. These operations would range from convolutional and fully-connected layers to Transformer blocks, according to Hongxiang Fan (2022). EdgeAI-Hub hardware should be optimised for a wider range of operations, i) support for multiple precisions, iii) support for sparsity and dynamic input length, iv) large multi-level memory for training with Direct Memory Access (DMA) support for zero-copy distributed machine learning, v) virtualisation for fast context switching and task preemption under multi-user tenancy, and vi) hardware-based app sandboxing with TEE for secure processing of sensitive data. This is in contrast to the SoCs that are currently used in smartphones. When it comes to the network, we foresee the EdgeAI-Hub being able to enable simultaneous communication over many interfaces. This would allow it to accommodate a variety of devices, as well as provide load-balancing and greater throughput. In addition to mesh networking using proxy repeaters for densely covering the deployment, Multiple-Input Multiple-Output (MIMO) over multiple antennas might be used to boost the capacity of communication.

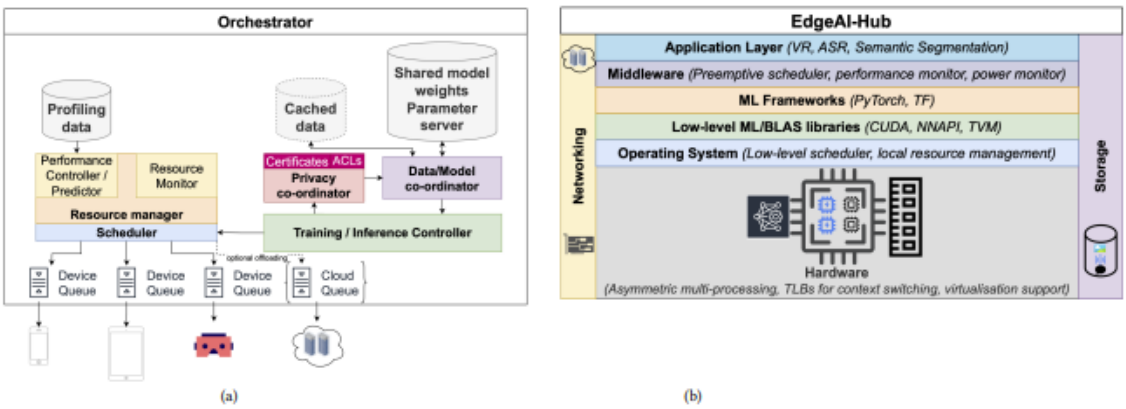

Figure 5: both the reference design for the orchestrator (4a) and the reference stack for the EdgeAI-Hub (4b).

setting up. Lastly, it is important that device discovery and handshaking be able to take place over any channel that is supported, such as Bluetooth Low Energy (BLE), Near Field Communication (NFC), or Ultra Wide Band (UWB). It is also possible that the EdgeAI-Hub's technology might be future-proofed through the use of replaceable components, and that this could sacrifice energy efficiency in exchange for more bandwidth. In artificial intelligence, data are considered to be first-class citizens; hence, storage becomes a crucial factor. We present a storage solution that is hierarchical and has quick caching for iterative processes such as model training or model/context sharing. Additionally, we suggest the use of conventional drives for long-term persistence purpose. When it comes to protecting the confidentiality and security of data, it is essential to implement hardware-level encryption and user management that includes support for Access Control Lists (ACL) from the filesystem. Last but not least, redundancy might be provided by means of hardware (for example, RAID) or distributed replication. As we move up the hierarchy, the operating system (OS) is accountable for the management of local resources and the scheduling of projects. For the purpose of DNN execution, low-level machine learning compilers and BLAS libraries would be placed on top of the operating system, coupled with high-level machine learning framework interpreters. The two levels that make up the top of the stack are the middleware and application layers. In addition to being the orchestrator, preemptive scheduler, and performance monitor of both local and distant resources, the former department is in charge of coordinating the execution of activities that are deployed across several locations. After then, the application could be applicable to a variety of use cases that are discussed in the next section.

6. Future Research Directions

Despite the fact that this study sets the basis for an Edge-AI-powered framework for zero-latency consumer micro-segmentation, there are a number of chances to further expand and enhance the methodology. Building upon the work that has already been done, the following suggestions for future study areas are recommended:

6.1 Enhanced Model Compression Techniques

It is possible for future research to investigate more advanced methods of model compression, such as neural architecture search (NAS), quantization-aware training, and knowledge distillation, with the goal of deploying models that are even more effective on ultra-low-power edge devices. The Segmentation of Context-Aware and Multimodal Behaviour A considerable improvement in consumer profiling might be achieved by including new data streams into the segmentation process. These data streams could include speech, picture, or location data. There is a possibility that research in multimodal learning at the edge might result in segmentation models that are more holistic and adaptable. Systems that are both Adaptive and Self-Learning A potential field is the development of edge-AI systems that are capable of autonomously adapting to changing client behaviours in real time without the need for centralised retraining. It is possible that in the future, research may concentrate on strategies for online learning, continuous learning, and meta-learning that can be used in non-stationary circumstances Sarker, I. H. (2021).

6.2 Secure Federated Learning Enhancements

For the purpose of ensuring that end-to-end data security is maintained during the training process, future research should study the possibility of integrating federated learning with privacy-enhancing technologies such as secure multiparty computing, differential privacy, and homomorphic encryption. Interoperability and standardisation of procedures For large-scale deployment, it will be necessary to construct edge frameworks that are vendor-agnostic and build across several platforms. In the future, research may lead to the creation of open standards and application programming interfaces (APIs) that enable interoperability across a variety of software and hardware ecosystems. Research on Human-in-the-Loop Systems might investigate the possibility of incorporating human input into the segmentation loop. This would be especially beneficial for applications that need ethical oversight, explainability, or domain-specific judgement, such as those in the healthcare or financial sectors. Scalability within Ecosystems That Are Federated The difficulties associated with maintaining model convergence and synchronisation increase in proportion to the number of edge devices that are participating. For the purpose of ensuring robustness and efficiency, it is recommended that future research investigate scalable aggregation algorithms, incentive

mechanisms, and decentralised learning topologies Zhang, C., Patras, P., & Haddadi, H.

6.3 Cross-Domain Application Studies

It is possible to get significant insights about the flexibility and effect of this Edge-AI architecture across a variety of domains by expanding its application to sectors other than retail and digital services. These sectors include education, healthcare, agriculture, and public transportation, among others. Optimisation of Energy-Aware Artificial Intelligence at the Edge It will be essential to do research on energy-efficient inference and training procedures in order to accommodate deployment in contexts with limited battery capacity. When it comes to scheduling, offloading methods, and dynamic power management, this comprises everything. Considerations That Are Both Ethical And Regulatory The ethical implications of real-time micro-segmentation should be investigated in further study. These implications should include the reduction of bias, the promotion of transparency, and the adherence to ever-changing worldwide data protection standards. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019).

7. Conclusion

The desire for real-time, personalised client engagement has been the driving force behind the progression of classic segmentation approaches towards micro-segmentation tactics that are more granular and dynamic. On the other hand, cloud-based systems are subject to substantial limits in terms of latency, privacy, bandwidth, and scalability, which hinders their capacity to deliver real-time business intelligence at a large scale. In this study, a unique Edge-AI architecture is proposed with the intention of enabling zero-latency consumer micro-segmentation. This is accomplished by processing behavioural and contextual data directly at the edge of the network. The system guarantees real-time responsiveness, increased data privacy, and decreased dependence on centralised infrastructures by deploying lightweight machine learning models and integrating federated learning. These features are achieved through the integration of federated learning. Our findings indicate that the Edge-AI technique is capable of achieving performance that is equivalent to or even greater to that of traditional approaches, while simultaneously lowering the amount of time required for inference and limiting the dangers associated with privacy. Because it is flexible across domains such as retail, banking, and digital services, the architecture that has been developed is a viable option for organisations that are looking to create hyper-personalized, real-time customer experiences. Furthermore, this work makes a contribution to the expanding corpus of research on edge intelligence by providing an outline of a micro-segmentation pipeline that is scalable, responsive to context, and conscious of privacy concerns. Additionally, this pipeline is capable of operating under real-world resource restrictions. When this framework is successful, it paves the way for more extensive applications of artificial intelligence at the edge of the network in other areas of real-time analytics and decision-making. As we look to the future, it will be vital to make more improvements in the areas of model compression, federated optimisation, and ethical regulation of artificial intelligence in order to refine and expand this framework. The next generation of customer intelligence systems will be significantly influenced by Edge-AI for micro-segmentation, which will play a crucial part in the process of enterprises continuing their transition towards decentralised computing paradigms.

Reference

[1] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Ramage, D. (2019). Towards federated learning at scale: System design. *Proceedings of the 2nd SysML Conference*.

[2] Chen, J., Ran, X., & Khoshgoftaar, T. M. (2019). Deep learning for edge computing: Review, opportunities and challenges. *Journal of Big Data*, 6(1), 10. https://doi.org/10.1186/s40537-019-0176-7

[3] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), 637-646. https://doi.org/10.1109/JIOT.2016.2579198

[4] Prakash Mallya, "High-Performance Edge Computing," Fortune India, Oct 21, 2019. [Online]. Available: https://www.fortuneindia.com/opinion/high-performance-edge-computing/103693

[5] Albert Zomaya, "Keynote 2: Resource Management in Edge Computing: Opportunities and Open Issues," 2019 IEEE Symposium on Computers and Communications (ISCC), 27 January 2024.

[Online]. Available: https://ieeexplore.ieee.org/document/8969601

[6] Chellammal Surianarayanan et al., "A Survey on Optimization Techniques for Edge Artificial Intelligence (AI)," Sensors 2023, 23(3), 1279, 22 January 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1279

[7] Identity Management Institute, "Edge Computing Security and Challenges." [Online]. Available: https://identitymanagementinstitute.org/edge-computing-security-and-challenges/

[8] U. Palani et al., "Edge Computing Based Autonomous Robot for Secured Industrial IoT," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 14 October 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9916016

[9] Codiant, "Edge Computing: The Next Generation of Innovation," June 12, 2023. [Online]. Available: https://codiant.com/blog/edge-computing-the-next-generation-of-innovation/

[10] Hongxiang Fan, Thomas Chau, Stylianos I. Venieris, Royson Lee, Alexandros Kouris, Wayne Luk, Nicholas D Lane, and Mohamed S. Abdelfattah. Adaptable Butterfly Accelerator for Attention-based NNs via Hardware and Algorithm Co-design. In International Symposium on Microarchitecture (MICRO), 2022.

[11] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

[12] Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8), 1738–1762. https://doi.org/10.1109/JPROC.2019.2918951