Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-753-6_1

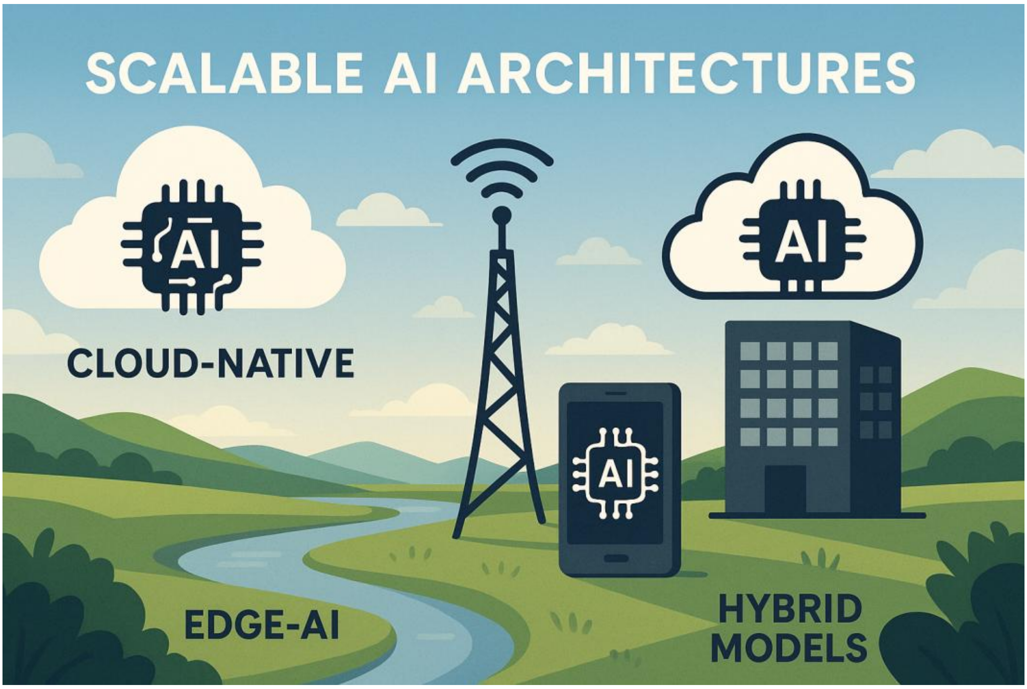
DeepScience
Open Access Books

Chapter 1: Scalable artificial intelligence architectures: Cloud-native, edge-AI, and hybrid models

Swarup Panda
SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have matured over the past two decades as decision-making technologies for systems and processes in varied domains [1-3]. AI and ML solutions are increasingly being moved from on-premises deployments to the Cloud, for multiple logical and practical reasons related to business agility, economical operations, performance at scale, availability, and security [2,4]. With the advent of the Internet of Things (IoT) and the increasing use of embedded or device-level intelligence for real-time decision-making and information filtering, there is a recent movement to define solutions that blend the Cloud with Edge devices. The goal is to leverage the benefits of both Cloud and Edge in a Hybrid architecture to solve specific business problems. From a research and education perspective, there are key questions that await answers: what is Cloud-Native AI Architectures? What is Cloud + Edge - AI Architectures? What are Hybrid Models? How do we design, code, test, deploy, and manage lifecycle for Cloud-native, Edge-AI, and Hybrid Models? What types of business problems are best solved using one of these architectures, and how do I know? What AI models work best in these environments? How do I put in Cloud, Edge, and Hybrid deployment best practices? To operationalize and automate these questions, what tools, techniques, and platforms do we need? This book attempts to answer these questions in its humble way, while remaining technology-agnostic wherever possible. The book takes the position that we will never be able to automate everything into a no-code AI solution [5-8]. We remain cognizant of the fact that building, deploying, and managing scalable AI solutions in Production is a challenging task that requires teamwork from multiple shared, but cross-skilled resources in Solution design, coding, testing, deployment, and management as part of a SLDC process involving tools and techniques from Software Engineering, DevSecOps, and MLOps.



2. Overview of AI Architectures

AI platforms typically use a digital infrastructure consisting of hardware such as chips, storage and data centres, combined with connectivity technologies such as the Internet, 5G and private networks, data and AI developer services such as MLOps, and application platforms specific to an industry or a region, such as robotics or language services [6,9]. Various features of this infrastructure define various cloud-native, edge AI, and hybrid models of AI. A focus on one or another model has an impact on the volumes, the types, or the speed of various AI services. In what follows, we consider the foundations of the various types of infrastructure, the features that define the major types of infrastructure, and then explore in more detail the cloud-native, edge-based, and hybrid types of AI. The specific infrastructure and service location options explored in this section, coupled with varying demand from customers, leads to the various types of cloud-native, edge AI, and hybrid models of AI discussed in the next three chapters.

The unique digital physics of AI leads to major differences between AI work and traditional applications like those driving big data [10-12]. Model and data parallelism, data-driven training, the availability of model pretraining and fine-tuning, and other features lead to phases of constant scale efficiency that permit cost-effective at-axis scaling of data and compute resources. When coupled with application-sensing infrastructure such as data services that help developers identify useful datasets, data marketplaces, MLOps and other AI cloud tools, the underlying infrastructure choices lead to unique design patterns in AI. Indeed, these design patterns differentiate AI architectures from those traditionally used for other types of business applications while also illustrating how AI patterns are derivatives of the traditional design archetypes.

3. Cloud-Native AI Architectures

The term "cloud native" has become synonymous with modern, scalable application development and deployment. Many well-known Internet services have adopted cloud-native architectures based on containers and microservices, and machine learning-based applications such as chatbots and search engines are inherently suited for cloud-native implementations [7,13-16]. Cloud-native refers to platforms that are purely built and run on the cloud without needing to rely on ground infrastructures in the form of data centers for hosting. Furthermore, cloud-native applications are designed, implemented, and deployed to take advantage of the specific characteristics of cloud services and solutions, such as big data processing, massively parallel computing or on-demand scalable computing resources. Cloud-native AI models are thus breaking away from traditional designs and constraints, and embrace and pursue all types of deployment on the cloud and with the cloud services.

The move to cloud-native AI is yielding significant benefits regardless of which type of cloud service and model are being used [2,17-19]. For example, the cloud enables AI developers to provision compute nodes with any specialized hardware such as graphic processing units or tensor processing units without needing to maintain these resources in-house. Furthermore, cloud providers offer several dedicated AI development toolkits, libraries, and frameworks that support the cloud environment and can significantly improve the productivity of AI development. Last but not least, resources provisioning in the cloud can be elastic, meaning that the provisioning can scale up or scale down seamlessly based on demand [3,20-23]. For organizations developing and deploying AI systems internally, the ability to provision a large number of compute nodes for a period of time, but relying on limited on-premise resources the rest of the time can lead to massive over-provisioning and thus savings in both capital and operational expenditures.

3.1. Definition and Characteristics

Cloud-Native architectures have emerged as a choice model for hosting and scaled-out services ever since the advent of fast Internet and the evolution of cloud computing as a model for hardware resource allocation of services. Such an implementation abstracts away the infrastructural overhead of hardware provisioning, deployment, and management, allowing service developers to focus on the service logic, while introducing several key features for automatic resource elasticity, service reliability, and availability. Vital to the model are Cloud APIs, available from Public Cloud service providers and various Private Cloud implementations that package not just Infrastructure as a Service, but also Platforms and various Software as Services to enable rapid Cloud-Native service deployment.

Cloud-Native services are implemented as microservices, using pre-packaged optimized functions, often Model Inference Compliance Libraries that accelerate the deployment of optimized complex mathematics on scalable Cloud hardware, and take advantage of automated orchestration and development tools that are designed to automate key aspects of the Deployment including load balancing, fail-over, scaling, reliability, and service optimization. Hardware support is provided through GPGPU/FPGA/TPU accelerated computing hardware, the high-bandwidth memory pooling/integrated architecture between CPU/GPU, high-bandwidth memory integration architecture, and other vendors, the use of Cache, Fusion and Body-Back Fusing Bandwidth techniques.

3.2. Benefits of Cloud-Native Models

Cloud-native AI architectures are built on a microservice-based design pattern that exploits the distributed, flexible, elastic, and on-demand computing, communication, and storage capabilities of modern cloud infrastructures while enabling a high level of parallelism and dynamic reconfiguration [9,24-26]. As highly domain-agnostic general-purpose deep neural network models, cloud-native AI models are scalable and efficient.

Despite successfully demonstrating superhuman performance and capabilities in sophisticated cognitive tasks for text, speech, vision, reinforcement learning, and multimodal domains, cloud-native AI models are yet to be adapted to new application contexts [27-29]. They are trained exponentiation faster and lead to a higher level of accuracy, performance, quality, and diversity when compared to small-customized, domain-specific, specialized AI models that are often optimized and finetuned for specific applications. For neural machine translation, the use of domain-specific, smaller customized models is known to deliver improvements compared to cloud-native models. However, it is costly and laborious to train a small-customized domain-specific model that needs to be regularly updated for newer domain data. As a result, organizations and businesses need to spend resources and costs to maintain multiple customized models that need to be relabeled and regularly retrained. For both financial documents and shortening the updating time of the model, the use of large cloud-native models is more efficient, requiring less time for retraining compared to the financial-domain custom models. Hence, considering the maintenance and retraining costs, using a cloud-native model is considerably cheaper than maintaining multiple customized ones.

3.3. Challenges in Implementation

Building AI systems that are entirely cloud native is not trivial, some questions remain open about the trade-offs that can be made with such a concentrated strategy. First, the cloud-native character of such systems may also reveal some serious weaknesses. The concept of cloud-native imposes high levels of distribution, fault tolerance, continuous evolution, disaggregation, and efficiency, but often neglects low latency and privacy. There are specific applications that need to be very low on latency or cannot afford to share their data with third-party infrastructure companies in order to train better analytics. Such concerns are especially strong in areas like the health sector, where sharing data is either limited by law or, in less strict countries, frowned upon.

Other features that we take for granted in non-AI systems, such as personalization, are also hard to conceive. Some personalization methods that work in non-AI systems like caching or content in databases can hardly work in AI systems. For example, performing a sequence-to-sequence AI constant transfer transformer on vast batches to group information flows for a specific user may not seem wise economically [30-32]. On the other hand, AI systems that are clothing personalized could be cloud deploying architecture with little to no specificity of the cloud-native paradigm, the cache, and the usual personal context data valid for such a specific user. But by the same token again, the requirement of a cloud AI learning system that finely tunes all the time is likely to be another economic fault with the deployment.

3.4. Case Studies

Although the world has many cloud-native products, few are recognized as cloud-native AI systems. Here, we look at some industry leaders.

Google Search has become a dominant means of finding and organizing information. Its proprietary algorithm is a cornerstone of the company's cloud-native AI system, making web-document ranking initially possible, and ultimately scalable. It famously ranks hundreds of billions of web pages. In addition to this algorithm, Google has incorporated many machine-learning components for personalization, formalization, high-quality content recognition, and monetization. The personal assistant, Google Assistant; the automatic captioning of videos; the app, which recognizes faces; the app, which translates text; and the service, which recommends routes, are also core features of the company's AI architecture, and their quality has been constantly improved as their respective neural backbones have evolved to more accurate, robust, and efficient architectures.

Facebook is known for its cloud-native social interaction platform. It is also known for its highly sophisticated AI system that recommends groups, finds friends' photos, and serves as an advertising engine. Facebook has developed AI technologies that leverage AI so that the company can scan through every produced content, including images, improve search, and make better personalization and monetization decisions. Facebook's AI Research lab is also a main contributor to a widely used deep-learning framework.

4. Edge-AI Architectures

We show that widening processing centrality or decentralizing execution to cost-effective managed edge devices for data analytical web services provides various compute performance, security, privacy and latency QA perks that AI needs. Although all computing can in principle go to the cloud or big data centers, any web-scale data analytic that generates high traffic volume in either direction, such as search, recommendation or user data learning from surveillance video, will benefit by distributing the process to the edge [9,33-35]. This is even greater when the analytic has a high service-rate request load but the server has limited compute acceleration boost, offering the edge model without as many latency penalties from slow transfer speed for small requests.

Edge-AI Models such as Cloudlet or fog computing work by also adding more microdata centers to the network. These are usually much smaller than big data centers, sometimes running on commercial load balancer clusters or simple, cost-effective appliances plugged into a wall outlet. These microdata centers are usually provisioned additionally to powerful edge networking devices like WiFi access points and 5G RANs [36-38]. These can pipeline any process with clear parallelism or batching across edge clusters, sending parts of the request, or dividing the request internally to use the edge device today, returning partial results or doing transmission to centralize the response. Models demanding high throughput but suffering delays due to back and forth transfer to the central cloud are ideal candidates to favor edge distribution. Sensors now powered by batteries and harvesting energy can make persistent Internet connectivity wirelessly free while running autonomously for unattended operations.

We also argue for minor data sharding or what we call micro-micro data centers at the third edge, embedded directly in the sensor performing the data collection or the other powerful device it may frequently transfer to [3,39-41]. Although requiring special hardware, research prototypes are feasible even for embedding at moore's law scale in silicon chip sensor cameras inspired by mobile device system on chip accelerators. Confined devices with very small computing resources connectable persistently wirelessly can perform dedicated, efficient pipelines for specialized ML workloads or micro-analytical processes that need to run on sensor devices for power conservation or immediate response are ideal candidates for tiny Edge Designs.

4.1. Definition and Characteristics

The explosive growth of interconnected smart devices is giving rise to unprecedented amounts of data that must be processed. Centralized cloud solutions are inefficient in processing the overwhelming amount of data involved. Hence, it has increasingly become the norm to deploy Artificial Intelligence models that partially offload computation on End Devices close to the User [36,42-44]. Edge Artificial Intelligence Models reside along the continuum from the Cloud, through Gateway Devices, down to the End Devices leveraging the amount of computation they can share and the resources they have. Edge-AI Models are increasingly considering the Bandwidth limitations and Latency requirements of different applications, deploying different model components at different locations to achieve a more Scalable Solution. Hardware Resources on Edge Devices are increasingly becoming more capable, Specialized Accelerators are deployed, and Hybrid Intelligence is being considered. Latency-sensitive applications need Data processed Close to the End-users, with Cloud or Nearby Gateway Solvers assisting in processing high-dimensional data. Latency-tolerant applications are suitable for partially processed data collaborating over Bandwidth-efficient Descriptive Statistics.

Edge AI Models potentially reduce the amount of Data that must be sent to the Cloud for Processing. Data from the End Device may be analyzed, summarized, or transformed using private or public models deployed on the Device, eliminating the need to transfer large amounts of Data to the Cloud for Processing [40,45]. The resulting light-weight Communication at Intermittent Intervals can greatly alleviate the Data Communication Bottleneck allowing for more Generative/Collaborative Model Components deployed on Cloud or Nearby Gateway Devices completing the Processing. Edge AI Models also increase the Levels of Data Privacy by keeping more of the Flow of Private Data localized. Even if Public Models are used to transform the Data on the End Devices, Details regarding the Input are not transferred to Cloud.

4.2. Benefits of Edge-AI Models

As shown in previous sections, there are many advantages in executing AI models at the edge by exploiting local computation power or at least hybrid schemes where computing is locally augmented by assistance from far more powerful computers residing in the cloud. Several edge-device characteristics need to be addressed by these models, including low latency, intermittent connectivity, data sovereignty/privacy, limited power, constrained bandwidth, resource-constrained devices, contextualization, and high scalability and automation. So far, typical edge-driven tasks are video/image analytics and speech recognition applications, mostly from the computer vision domain.

The increase in demand for smart, high-traffic user-heavy applications, guided by the current wave of AI generative models, calls for fast, cost-efficient models and applications. Yet today's cloud-centric services based on large-scale cloud infrastructure powering heavyweight models are struggling to keep up with that demand. We believe that a correct balance between these paradigms can solve this problem, further increasing model accessibility and decreasing ingress/egress costs. Moreover, the success of AI applications relies heavily on the ability to adapt to different scenarios, making application and training/serving costs prohibitive. Edge-AI models can alleviate some of these bottlenecks, such as the requirement of critical latency performance, since they operate locally and augment results with user data and other hints that compute-augmented cloud assistants provide. In other words, Edge-AI responding to local triggers can increase the overall responsiveness performance while offloading resource-intensive stages to the cloud. Moreover, by involving sensitive and personalized data, edge-centric operations can optimize access and computation on data locally generated and govern the sanitization of the information made available to potent but less responsive cloud services.

4.3. Challenges in Implementation

Despite the numerous advantages associated with self-sufficient system architectures and the accelerated development cycles they afford, there are still significant challenges to full adoption of Edge-AI approaches. Inherently minimal system specifications call for collaboration with suppliers to verify and document thermal, stability, and safety characteristics for hardware components under edge operating conditions. Factors such as minimizing the number of subsystems and physical packaging volume lead to tighter tolerance errors on electrical parameters than would be expected for a PC or workstation class system. Special consideration on the possible impedance and voltage gain deviations in circuits must be evaluated since these circuits are highly sensitive to temperature variations and can reflect highly nonlinear behaviors. Shortened product life cycles at the edge extend the challenge of increased reliability. Unlike components envisioned to reliably operate for years, edge devices operating in battery-powered mobiles or exposed to different user activities might experience far more stress cycles during their useful lives. Building Edge-AI designs that are power efficient yet remain within the average lifespan of consumer usage for battery powered devices is further exacerbated through considerations on battery memory effects.

Additional considerations complicating Edge-AI system designs arise from the continuing growth of heterogeneity in silicon capabilities to execute different components of the hybrid model. This gradual expansion of chip features makes dynamic execution flow management for real-time applications increasingly complicated. Achieving a higher level of model accuracy to best utilize the selection of heterogeneous platforms for lessons processed at the edge requires both faster converging algorithms capable of performing in transient states and improved confidence measures that can truly rate the inferential performance of the components involved.

4.4. Case Studies

Leveraging the entrepreneurial spirit of the business community and the availability of funding, the research center has built prototypes of intelligent video services based on different Edge-AI models with partners from different domains. Some of these prototypes have evolved into commercial video analytics products. By implementing these prototypes, various implementation challenges for different types of Edge-AI models were identified and insights were gathered into how to attain an optimal architecture for a particular application domain.

The first approach to Edge-AI implementation is to build a custom model based on a small Edge-AI platform and off-the-shelf hardware. As illustrated by the prototype called Air-Insightboard, which requires no prior training, a custom model could be applied when the supervised training data are too few to finetune a generic model. Air-Insightboard, a low-cost, custom model-based system prototype, was built in collaboration with a partner. Another custom model, called Jalebi Vision, was implemented at another partner in a different domain. Jalebi Vision, which required no prior training, sensed the dark web traffic of Jalebi, a popular Indian food item throughout the year, and was used to support the Jalebi client by estimating the demand during the upcoming week to enable suppliers (small local businesses) to plan the logistics accordingly, using corrected demand.

A second approach to implementing Edge-AI is deploying a model on a product platform, whereby the model performs a few but critical operations, while the model execution is offloaded via a wireless backhaul on higher capacity, more robust GPU-platforms located on the ground, near the operations with better internet connectivity.

5. Hybrid AI Architectures

The rapid evolution of Artificial Intelligence (AI) ushers in the era of Hybrid AI (HAI), characterized by and utilizing the interaction of different artificial components of intelligence at different time scales, as well as the interaction of those components with cognitive human intelligence, at multiple levels of abstraction in both the physical world and the digital world. We live in a Cyber-Physical Society (CPS), where humans and AI are co-creators and co-inhabitants of our common landscape, characterized by the fusion of relationships and connections within and between the physical world and the digital world. Such relations are already experienced at different levels of comfort, and increasingly will be experienced thanks to the scalable AI systems and solutions powered by Hybrid AIs. In the coming years, these systems will be enriched and augmented by more advanced expert-driven HAI systems, today termed Large Language Model (LLMs) with specializations, which will increasingly work seamlessly and cooperatively with our cooperative-oriented cognitive human abilities and our specialized expert knowledge. Cooperation and interaction between HAIs and humans is at the core of realizing the potential of a Hybrid Society (HS), empowered by trustworthy, explainable, human-aligned, cooperative-oriented Hybrid AIs. An HS is the virtual and physical use case in our daily activities and within all the dimensions of our communities, such as transport, education, social interactions, that are at the core of our civilization.

Our AI architectures should fuse Digital AI (DAI) models and solutions trained on natural, message-colored, logical, abstract, symbolic, structured, representation languages, fused with Specialized Raw Data Trained AI (DRT-AI) solutions trained on raw data models using pattern recognition and deep learning, that are at the core of today’s rapidly evolving AI products and systems, fueled by today’s Digital Gold Rush - Data! Such Hybrid Architectures (HAs) represent the logical further evolution of Modular, Coordinated, Architectures (MCAs). These proposed HAs will exploit the complementary advantages of DAI-Language approaches with the DRT-AI Neural Network approaches, at the core of today’s most advanced AI-powered enterprise systems and products.

5.1. Definition and Characteristics

Hybrid artificial intelligence (AI) architectures employ synergetic combinations of two or more AI base components or methodologies, heterogeneous or homogeneous ones, to realize added value in accomplishing intricate tasks beyond the performance of any specialized model. In the recent AI growth wave, hybrid architectures have taken prominence by judiciously tying advanced deep learning neural networks with conventional symbolic AI, knowledge graphs or knowledge bases, traditional machine learning classifiers, expert systems, and probabilistic graphical models including Bayesian networks, Kalman filters, or Hidden Markov models. The flexibility of designing hybrid architectures stems from the unique capabilities of each base model to tackle diverse, intricate tasks, confirmable also thanks to their historical maturation.

For instance, symbolic AI and neural networks are distinctive in their behavior; the former reason and explain well and prove generality, but require substantial human labor to codify their logical rules and knowledge explicitly. The latter learn solely from data, in a bottom-up fashion, focusing on recognition, recommendation, and regression for directly observable variables and states, and via tedious backpropagation and adjustment of millions of parameters centralizing the learning task. Nonetheless, they perform weakly in generalization and explainability. Moreover, a compelling challenge is the data-hungry nature of neural networks that need thousands of labelled data samples while some AI tasks lack sufficient reliable data. Thus, to solve intricate application problems, a natural option is to increment the strength and capabilities of the classical AI models in natural language processing, computer vision, and all areas of inference and classification by forming hybrid architectures; heuristically, to couple the cogitative logic symbolic structures and the real world statistics modeled with neural networks to leverage all the distinctive merits.

5.2. Benefits of Hybrid Models

Despite scalability being one of the most important benefits of cloud-native architectures, recent applications demonstrate that there are many cases where a hybrid model can offer significant advantages. For example, in the area of personal assistance, what could appear to be exclusively cloud-based applications have gone hybrid: even when most of the processing such as natural language comprehension or multi-modal understanding occur in the cloud, the timely delivery of outputs heavily depends on completing the process flow on the device. Consider a personal assistant device such as Google Nest, HomePod or Amazon Echo. These devices wake up via a voice trigger, process the incoming request, and then submit it to the cloud for natural language understanding.

There is a full advantage of doing these operations in the cloud, where data from a large set of users is used to create the word-spotting model. But the hybrid approach should be implemented as the smart device must also be able to complete the request elicited by users. For example, upon hearing the word trigger and after asking a question such as “What’s the weather like in San Francisco?” the device should immediately follow with “According to weather, it is foggy at 60 degrees”, just as a person would when completing that request. This requires an appropriate model to load on the device to ask and produce the output quickly, conservatively and with edge-AI engineering.

5.3. Challenges in Implementation

Hybrid AI models and architectures address the complementing limitations of cloud-native and edge-native AI approaches by merging the cloud and edge capabilities, resources, and services. Hybrid AI systems will however achieve their goal if they manage to overcome the challenges posed by the complexity and heterogeneity of the cloud and edge resources, and service layers. Among the challenges that both industry and research face when dealing with hybrid AI models we can enumerate: (1) Wide spectrum definition: A seamless integration of cloud-native and edge-native AI services may need to handle highly heterogeneous models. But what kind of models should integrate to achieve a mutually beneficial collaboration, and how? Should we add small, lower-impact edge parts of bigger impactful models, in a modular fashion, to cloud-native services instead of building smaller, distributed models for different edge tasks? (2) Open challenges and problems for partitioning models and architectures: After answering the prior question, what techniques can be leveraged to partition models and designs at the architectural level to enable distributed inference with major gains? Nowadays, partitioning distributed ML models such as NLP transformers and large Vision models is a real challenge, as partitioning procedures need to account for a plethora of design considerations when it comes to constraints to be put to the edge inference scenario. These design considerations include security, privacy, availability, fault tolerance, responsiveness, real-time requirements, and cost considerations either for deployment or inference on the heterogeneous edge infrastructure. What kind of architectural-level knowledge can we integrate, either at the application or network level, to achieve hybrid AI partitioning solutions?

5.4. Case Studies

Although some may think that hybrid models are a recent development, they are rather common in practice. Publications and literature mostly address either the cloud-nativeness or the local AI capabilities. Very few papers address hybrid capabilities from an architectural solution perspective. This section presents some real practical AI implementations, which integrate both cloud-native components along with edge-AI capabilities.

Many robotic systems perform AI tasks that rely on both cloud-based resources and onboard resources. The cloud is normally used for data storage, training, model initialization, and sync; while, the edge-based modules are used for inference and action commands. A commercial example of such hybrid approach is demonstrated by the scale of operation of cloud and onboard AI resources in the development of more than 3000 autonomous robotic arms.

The AI capabilities of such robotic arms make it suitable not just for industrial applications but also for general-purpose applications as well. It can be trained or programmed for such essential but intellectual-heavy tasks as robotic-assisted finance for business and account summaries or reports, warehouse operations, 3-D scanning design and manufacturing operations or prep, automobile service repairs or design patterns, security reasoning for tourists and customer service, hotel management service during seasons; etc. All such tasks can be implemented using a cloud-based architecture, with the inefficient and high-bandwidth usage state stage being implemented by local edge-AI modules.

6. Comparative Analysis of AI Architectures

The various AI architecture models have been facilitated and optimized for different use cases. In this section, we will conduct a comparative analysis of the various AI architecture models, focusing on the performance of the architecture models, as well as their operating cost and scalability. The performance will include inference delay, throughput, memory usage, and model size. The cost analysis will investigate both the economics of training the AI models on the central cloud as well as the cost of performing the inference on the AI deployment. We consider three aspects of costs, namely the monetary cost, the resource cost, and the environmental cost. Scalability is one of the main design criteria for AI models and algorithms, addressing the capabilities of handling larger data, more complex models, or meeting more stringent requirements on model performance. In this section, we investigate the implications of the design and deployment architecture on the scalability of the AI models and algorithms. The comparative analysis will provide a systematic review of the AI architecture and highlight the design tradeoffs for AI practitioners to consider when selecting the appropriate architecture for their application. In addition, it can help AI architecture researchers discover architectural improvements that address the gaps in this exploration, for example performance versus cost, or single model versus ensemble of models. The end goal is to provide AI researchers and practitioners with the knowledge to make informed decisions on their choice of infrastructure and architectures, and to systematically close the performance, cost and scalability gaps for their real-world AI applications via conscious investment in talent or solutions.

6.1. Performance Metrics

When designing and deploying AI applications, it is critical to assess the performance of these applications as a function of the platform used. Performance benchmarks are necessary to measure the end-to-end latency and throughput to help system architects make the right design choices. Ideally, benchmarks that capture the real workload of an AI service should be used to obtain the most accurate performance. However, due to several factors, such benchmarks are not always available. The model used, the timing mechanism used in the benchmarks, the controllers used in the AI service, the workload being executed, and the resources allocated for the inference execution affect the latencies reported by an AI service. Comparing inference latencies reported by different benchmarks. Long type AI workloads have latencies measured in tens of seconds, while mid-type AI workloads have latencies measured in hundreds of milliseconds, and short type AI workloads have latencies measured in seconds.

The most common AI benchmark selects AI workloads as the most impactful representative workloads. Inference latency for workloads such as image classification and object detection for computer vision, language translation, language modeling, document ranking for natural language processing, speech-to-text recognition for automatic speech recognition, recommendation engines for recommendation systems, and audio-visual recognition for multimodal AI are measured. Latency comparison of datacenter AI platforms. The workloads are classified into long, mid, and short latencies depending on the complexity of the AI inference. The workloads with shorter latencies are intended for execution in the datacenters while the workloads with longer latencies are intended for edge or backend execution. For the short and mid-type AI workloads and less, it is important for the latency of the workload to approach the performance target to meet AI goals.

6.2. Cost Analysis

The price of deploying hardware and software systems in the cloud is one of the major issues faced by organizations transitioning to cloud solutions. Given the variety of substantial initial financial and additional operational efforts needed to set up the service in a cloud environment, the goal for a successful service is to have a lower operational cost on such configurations, compared to on-premise ones. Therefore we will estimate what portion of the service is implemented at the various locations, how much the tariffs are for such components and what usage is expected over time.

The sampling frequency determines how many requests per second are expected, while there is a small delay for how long the model expects to work on a request. Then, based on the pricing of the active components of the hybrid FaaS, we can estimate the overall cost and, subsequently, the difference for the other offered solutions in terms of opportunities. The process can be repeated for the residential task, knowing that there the edge-device must be active continuously to be always-on. Also for portfolio applications, we have similar assumptions to the one applied for the local environment. Here we could consider that bigger requests are only processed in the cloud environment, with the other locations active for less time and processing various daily activities for each smart home. We summarize here the components that impact the final cost. How much of the functional architecture of the solution is hosted in the cloud and mobile part, and what is the downtime in such locations can also vary significantly from party to party, imposing both management for setup and control at the local level.

In such cases, we either need to inspect them periodically or verify from time to time the local component in action over a long time. We need to check the edge-device and how long it is on, but also know when it is active specifically processing a request. For such components we can obtain general behavior for the model, measure how long something is done and predicting how much other tasks expect otherwise do some checking. Such a periodic approach can be performed also while checking the cloud environment's request processing.

6.3. Scalability Considerations

All business-oriented software and systems must be designed considering their expected growth. AI systems have normal growth patterns of backend systems, but they also have an additional dimension, the data size and complexity, which is proportional to the expected growth in this domain. As the company's strategy matures in relation to AI, not only are solutions created to solve specific business needs, but also the amount of data that need to be processed to extract patterns and perform predictions becomes larger. This increase of data could rest on the specific solution for a company, or generally be based on a domain; for example, within a corporation there could be different sales, production, marketing, and finance systems, and all these systems as they expand, connect to external companies or entities for more consolidated analyses, they need to rely on bigger query processing, data management and AI model validations and predictions. In the case of a specific domain in the industry, the historical data in relation to all companies in this specific domain, while it does not grow as fast, a company collecting and analyzing its own amount of data becomes complex enough that the data expansion has effects on AI solutions.

There are basically two types of scaling techniques: scale up and scale out. Scale up is based on vertical scaling, these techniques rely on large servers that are able to process and store data close to AI solutions. Scale out is horizontal scaling, these techniques rely on several smaller servers that can be distributed geographically where the data resides. This local data processing allows the data to scale considering the company as well as domain type, and it basically does the same as the edge processing, only in a more centralized solution. However, cache storage is often needed to minimize the data exchange between local servers and central servers that execute the actual models.

7. Security and Privacy in AI Architectures

AI systems pose unprecedented challenges to privacy and security for technology innovators, corporations, and governments. AI systems rely on huge troves of proprietary, sensitive, and user-generated data, such as interactions with chatbots, visual inputs, and social media interactions, to act as intelligent proxies on behalf of users. A single AI model can be a trojan horse, containing security and privacy vulnerabilities for an entire class of systems. Various recent attacks on natural language processing systems have shown that adversaries are able to extract sensitive private information, such as user passwords, addresses, credit card numbers, and encryption keys, that have been disclosed by users in conversations with chatbots. Models trained using data from several user sessions also face a threat of backdoor attacks, where the attacker infects the model to secretly disclose sensitive information when the model sees a certain trigger string, while appearing to produce normal behavior otherwise.

Both cloud-native and edge-AI architectures face unique security and privacy challenges. The cloud needs to collect and query data from many users, and to be able to train systems, such that no sensitive user information is disclosed in training. Edge devices must maintain the user's privacy, without access to the global training datasets. Solutions must be pragmatic. AI is based on large datasets, and effective solutions must recognize that when in use by a user, the AI system is a reflection of that user. Mitigation approaches also must ensure that model performance is not severely reduced by privacy protections. For sensitive applications such as chatbots, mitigations could include employing the architecture such that the user acts as a human-in-the-loop, and uses shared control on the interaction protocol with the AI system. This could entail predefining keywords at the start of the session. Users must periodically review and update the keywords to prevent the model from disclosing sensitive information.

7.1. Threats and Vulnerabilities

To date, secure deployment is hardly contemplated in the AI development process, ranging from data collection, labeling, and model optimization to inference evaluation. In the particular area of cloud-native applications, robust and secure deployments are relevant not only for ML operators, but also ML developers and other stakeholders. Developmentwise, replicability and trust come into play, pointing at the need for tools ensuring secure experimentation of ML frameworks. The need for trust and secure experimentation raises questions of adversaries being present during the development, validation, and training stages of ML. An example of a potential attack vector at the data manipulation level is openly accessible data-labeling services, which can be exploited to introduce erroneous labels. To counter-act erroneous datasets, detecting poison data during validation or curating data-point prioritization for better and more reliable model performance are strategies suggested.

In addition to data and its handling, exploiting vulnerable frameworks used to build, validate, train, and optimize ML models also represent a real threat. More general weaknesses of software supply chains further raise the stakes, as building blocks are likely to rely on shared, publicly available underlying libraries or resources. To provide developers with insight into potential attacks at the model optimization stage, while improving model validation and training, a survey analyzing adversarial ML highlighted recognized model threats and generic defenses regarding the respective threats. Also at the model level, hidden threats when dealing with federated learning, which allows different clients to train models collaboratively without sharing sensitive data, were uncovered.

7.2. Mitigation Strategies

Due to the presence of deep learning frameworks, model access APIs, and hardware software optimization packages, the level of effort required in executing model and data poisoning attacks is minimal and the required skills set is not significant. Thus, it is necessary to invest in the construction of devices, models, and AI systems that are resistant to attacks. Model poisoning can be mitigated by hardening the training process using ensemble learning and secure training frameworks. Distributed learning allows incorporating diverse and independent datasets that can reduce the likelihood of manipulation by adversaries. Model/data inspection can also be used at certain training checkpoints to clean up the models and restore to earlier temporal states. By using cryptographic primitives during training or inference phases, the risk of bias poisoning is made negligible. These cryptographic primitives can also be generalized to advanced homomorphic encryption systems.

Most of the AI inference tools and pipelines applying ML/DL models are still static mappings from input to output. The predictions are usually binary-decided or for the likelihood of certain standard classes in CV, NLP, and ML. No uncertainty assessments are provided and such assessments are essential for sensible automated-decisions during the deployment of AI systems. A possible route of research in making AI systems resilient to data sampling vulnerabilities and adversarial-attacks is to augment the present deterministic model-outputs with uncertainty, over-distribution variance, and confidence quantification in the output-space. This will allow making risk-aware decisions by the downstream AI application services and business owners when deploying for real decision tasks.

8. Future Trends in AI Architectures

In this chapter, we consider which opportunities in AI technologies may awaken in the near future, and what sweat will be spent on implementation of the transition of AI for companies to cloud-native, edge and hybrid architectures. We will philosophize a little about what dilemmas AI software closing provides for software developers, and what bottlenecks await the pioneer of the transition to distributed types of AI architecture. We will brainstorm a bit about what needs to be done now in applied AI architecture engineering to create the necessary conditions for a successful software deployment.

We note right away that in this discussion, we are more likely to adhere to supporter rather than critic positions towards such mass use of AI technologies for various businesses. This predisposition arises from the fact that over the past years, switching to cloud-native architectures has created a number of complex, intricate and low-obvious problems. A number of companies developing systems and software for special task types have opportunistically switched to simplifying such transition to cloud-native architectures by offering tools for so-called cloudification. AI-tasks have a number of events similar to IT-tasks but have a number of differences. Thus, some companies have made enough investments and these problems are stabilized or successfully solved. However, many other companies face bottlenecks waiting for the needed AI-solutions to find their AI-driver and pioneer the AI-software project toward successful solution.

8.1. Emerging Technologies

As AI is transitioning from research to the mainstream, we are seeing massive investments in building supportive infrastructure and service stacks. There is also a sub-set of infrastructure technology innovation that will deliver the next-generation AI capabilities addressing the following critical deficiencies in the current stack: accessible, capable, and secure infrastructure geared into managing large data sets and transferring them for training; tools to simplify large data set engineering and cleaning; AI models with enhanced reasoning capabilities that can utilize external knowledge; efficient training and inference on custom-designed hardware; runtime and monitoring tools to track complex large-model jobs; and reasoning trusted AI stacks capable of integrating modular ontology and other logical reasoning stacks.

The primary emerging technology directions we see include: native support within the cloud-native IaaS offering focused on stateful, secure infrastructure support, addressing large storage and data transfer needs and built-in support for alternative compute units; cloud service products for simplifying large dataset engineering, such as versioned scraping, sampling, augmenting, and key-structure-enhancing; crowd services for annotating and explaining datasets; modular reasoning stacks for probabilistic logic combination; custom-designed, multi-architecture hardware built around AI reasoning workloads; edge platforms and devices capable of localized training and fine-tuning for support of on-device reasoning; and modular AI tools for comprehensive monitoring and management of complex AI model pipelines.

8.2. Predictions for Development

Various models and architectures will dominate at different scopes/timelines in the future; we make the following predictions for which models will be used when.

• At a very small scale (single-device AI, L × W pixels), classic models will still be the state of the art for the next 15 years. These classic models are efficient and fast, have small memory footprint, and can be fine-tuned to current hardware ecosystems. We also need to develop better and better models over time for such small scales due to various reasons. Hence, we predict such classic models exist for the next 15 years, especially for using edge devices in small models. More advanced models will still be too heavy and consume too much resources.

• For small scale (medium-device AI, L × W pixels), we will see transformer models as the state of the art in the next 15 years. However, at this scale, such models are mostly used for text/image generation and face recognition-type applications. Concerning similar questions for such models, classic models will still exist for similar reasons as we stated in the previous point. For this case, such models can also be optimized for speed/memory limitation since fine-tuning is relatively efficient and common.

• For medium scale (medium-device and distributed-device AI, L × W pixels), computer vision transformers will be predominant patterns for video classification/action detection, etc. Similar to other small/micro models, classic models will still be around for why we discussed in other areas.

9. Conclusion

Cloud-native AI, edge AI, and hybrid cloud-edge AI will accelerate the adoption of AI, making it faster and cheaper to build and deploy machine learning and deep learning applications. Cloud-native AI models will democratize AI use to everyday developers without needing PhDs in machine learning or deep learning to build models in a matter of hours, instead of weeks or months. The use of pre-trained cloud-native models and APIs can enable the embedding of NLP and image processing capabilities on cloud-enabled applications without the need to build and train special purpose models. Hybrid cloud-edge AI models deploy more sophisticated workloads where they can take advantage of the cloud environment. Combining cloud-based computing cycles with smaller amounts of local-edge computing cycles can accelerate the time to available AI results compared to exclusive cloud-based or edge-based AI workloads.

AI is an incredibly powerful tool expected to transform industries across the globe, from Finance to Healthcare, to Logistics, Supply Chain, and Shipping. But there are challenges associated with this tremendous transformative tool. Any successful business doing business at scale cannot afford to have downtime, and expects near 100% reliability, performance, and predictable results. Most prebuilt AI models are black boxes; it's difficult for users to anticipate how the model will respond to edge cases, and it has proven difficult for users to debug models that return unexpected or inappropriate results. Models tend to work well on training and test datasets that match or approximate modeled scenarios or cases, but humans (and systems) are clever, and edge cases can result in unexpected or inappropriate results. Innovators pursuing this massive opportunity must be enabled to speak the language of required AI product use, without getting buried in the complexities of lower level model architecture and engineering.

References:

[1] Sanz JL, Zhu Y. Toward scalable artificial intelligence in finance. In2021 IEEE International Conference on Services Computing (SCC) 2021 Sep 5 (pp. 460-469). IEEE.

[2] Haefner N, Parida V, Gassmann O, Wincent J. Implementing and scaling artificial intelligence: A review, framework, and research agenda. Technological Forecasting and Social Change. 2023 Dec 1;197:122878.

[3] Sai S, Chamola V, Choo KK, Sikdar B, Rodrigues JJ. Confluence of blockchain and artificial intelligence technologies for secure and scalable healthcare solutions: A review. IEEE Internet of Things Journal. 2022 Dec 29;10(7):5873-97.

[4] Moro-Visconti R. Artificial Intelligence-Driven Digital Scalability and Growth Options. InArtificial Intelligence Valuation: The Impact on Automation, BioTech, ChatBots, FinTech, B2B2C, and Other Industries 2024 Jun 2 (pp. 131-204). Cham: Springer Nature Switzerland.

[5] Oikonomou EK, Khera R. Designing medical artificial intelligence systems for global use: focus on interoperability, scalability, and accessibility. Hellenic Journal of Cardiology. 2025 Jan 1;81:9-17.

[6] Sayed-Mouchaweh M, Sayed-Mouchaweh, James. Artificial Intelligence Techniques for a Scalable Energy Transition. Springer International Publishing; 2020.

[7] Govea J, Ocampo Edye E, Revelo-Tapia S, Villegas-Ch W. Optimization and scalability of educational platforms: Integration of artificial intelligence and cloud computing. Computers. 2023 Nov 1;12(11):223.

[8] Hammad A, Abu-Zaid R. Applications of AI in decentralized computing systems: harnessing artificial intelligence for enhanced scalability, efficiency, and autonomous decision-making in distributed architectures. Applied Research in Artificial Intelligence and Cloud Computing. 2024;7(6):161-87.

[9] Pazho AD, Neff C, Noghre GA, Ardabili BR, Yao S, Baharani M, Tabkhi H. Ancilia: Scalable intelligent video surveillance for the artificial intelligence of things. IEEE Internet of Things Journal. 2023 Mar 31;10(17):14940-51.

[10] Sakly H, Guetari R, Kraiem N, editors. Scalable Artificial Intelligence for Healthcare: Advancing AI Solutions for Global Health Challenges. CRC Press; 2025 May 6.

[11] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.

[12] Bano S, Tonellotto N, Cassarà P, Gotta A. Artificial intelligence of things at the edge: Scalable and efficient distributed learning for massive scenarios. Computer Communications. 2023 May 1;205:45-57.

[13] Mishra A. Scalable AI and Design Patterns: Design, Develop, and Deploy Scalable AI Solutions. Springer Nature; 2024 Mar 11.

[14] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.

[15] Abisoye A, Akerele JI. A scalable and impactful model for harnessing artificial intelligence and cybersecurity to revolutionize workforce development and empower marginalized youth. International Journal of Multidisciplinary Research and Growth Evaluation. 2022 Jan;3(1):714-9.

[16] Raman R, Buddhi D, Lakhera G, Gupta Z, Joshi A, Saini D. An investigation on the role of artificial intelligence in scalable visual data analytics. In2023 International Conference on Artificial Intelligence and Smart Communication (AISC) 2023 Jan 27 (pp. 666-670). IEEE.

[17] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.

[18] Newton C, Singleton J, Copland C, Kitchen S, Hudack J. Scalability in modeling and simulation systems for multi-agent, AI, and machine learning applications. InArtificial Intelligence and Machine Learning for Multi-Domain Operations Applications III 2021 Apr 12 (Vol. 11746, pp. 534-552). SPIE.

[19] Bestelmeyer BT, Marcillo G, McCord SE, Mirsky S, Moglen G, Neven LG, Peters D, Sohoulande C, Wakie T. Scaling up agricultural research with artificial intelligence. IT Professional. 2020 May 21;22(3):33-8.

[20] Meir Y, Sardi S, Hodassman S, Kisos K, Ben-Noam I, Goldental A, Kanter I. Power-law scaling to assist with key challenges in artificial intelligence. Scientific reports. 2020 Nov 12;10(1):19628.

[21] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.

[22] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.

[23] Shlezinger N, Ma M, Lavi O, Nguyen NT, Eldar YC, Juntti M. Artificial intelligence-empowered hybrid multiple-input/multiple-output beamforming: Learning to optimize for high-throughput scalable MIMO. IEEE Vehicular Technology Magazine. 2024 May 20;19(3):58-67.

[24] Samuel O, Javaid N, Alghamdi TA, Kumar N. Towards sustainable smart cities: A secure and scalable trading system for residential homes using blockchain and artificial intelligence. Sustainable Cities and Society. 2022 Jan 1;76:103371.

[25] Villegas-Ch W, Govea J, Gurierrez R, Mera-Navarrete A. Optimizing security in IoT ecosystems using hybrid artificial intelligence and blockchain models: a scalable and efficient approach for threat detection. IEEE Access. 2025 Jan 22.

[26] Mungoli N. Scalable, distributed AI frameworks: leveraging cloud computing for enhanced deep learning performance and efficiency. arXiv preprint arXiv:2304.13738. 2023 Apr 26.

[27] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.

[28] Cheetham AK, Seshadri R. Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery. Chemistry of Materials. 2024 Apr 8;36(8):3490-5.

[29] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.

[30] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.

[31] DeCost BL, Hattrick-Simpers JR, Trautt Z, Kusne AG, Campo E, Green ML. Scientific AI in materials science: a path to a sustainable and scalable paradigm. Machine learning: science and technology. 2020 Jul 14;1(3):033001.

[32] Klamma R, de Lange P, Neumann AT, Hensen B, Kravcik M, Wang X, Kuzilek J. Scaling mentoring support with distributed artificial intelligence. InInternational Conference on Intelligent Tutoring Systems 2020 Jun 3 (pp. 38-44). Cham: Springer International Publishing.

[33] Otaigbe I. Scaling up artificial intelligence to curb infectious diseases in Africa. Frontiers in Digital Health. 2022 Oct 21;4:1030427.

[34] Dasawat SS, Sharma S. Cyber security integration with smart new age sustainable startup business, risk management, automation and scaling system for entrepreneurs: An artificial intelligence approach. In2023 7th international conference on intelligent computing and control systems (ICICCS) 2023 May 17 (pp. 1357-1363). IEEE.

[35] Peteiro-Barral D, Guijarro-Berdiñas B. A study on the scalability of artificial neural networks training algorithms using multiple-criteria decision-making methods. InInternational Conference on Artificial Intelligence and Soft Computing 2013 Jun 9 (pp. 162-173). Berlin, Heidelberg: Springer Berlin Heidelberg.

[36] Kuguoglu BK, van der Voort H, Janssen M. The giant leap for smart cities: Scaling up smart city artificial intelligence of things (AIoT) initiatives. Sustainability. 2021 Nov 7;13(21):12295.

[37] Gowda D, Chaithra SM, Gujar SS, Shaikh SF, Ingole BS, Reddy NS. Scalable ai solutions for iot-based healthcare systems using cloud platforms. In2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) 2024 Oct 3 (pp. 156-162). IEEE.

[38] Awan MZ, Jadoon KK, Masood A. Scalable and effective artificial intelligence for multivariate radar environment. Engineering Applications of Artificial Intelligence. 2023 Oct 1;125:106680.

[39] Landin M. Artificial intelligence tools for scaling up of high shear wet granulation process. Journal of Pharmaceutical Sciences. 2017 Jan 1;106(1):273-7.

[40] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.

[41] Mocanu DC, Mocanu E, Stone P, Nguyen PH, Gibescu M, Liotta A. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nature communications. 2018 Jun 19;9(1):2383.

[42] Blanco L, Kukliński S, Zeydan E, Rezazadeh F, Chawla A, Zanzi L, Devoti F, Kolakowski R, Vlahodimitropoulou V, Chochliouros I, Bosneag AM. Ai-driven framework for scalable management of network slices. IEEE Communications Magazine. 2023 Nov 23;61(11):216-22.

[43] Sadek AH, Mostafa MK. Preparation of nano zero-valent aluminum for one-step removal of methylene blue from aqueous solutions: cost analysis for scaling-up and artificial intelligence. Applied Water Science. 2023 Feb;13(2):34.

[44] Cohen RY, Kovacheva VP. A methodology for a scalable, collaborative, and resource-efficient platform, MERLIN, to facilitate healthcare AI research. IEEE journal of biomedical and health informatics. 2023 Mar 20;27(6):3014-25.

[45] Adelodun AB, Ogundokun RO, Yekini AO, Awotunde JB, Timothy CC. Explainable artificial intelligence with scaling techniques to classify breast cancer images. InExplainable

Machine Learning for Multimedia Based Healthcare Applications 2023 Sep 9 (pp. 99-137). Cham: Springer International Publishing.