Motivation, Definition, Application and the Future of Edge Artificial Intelligence

[1]Anandakumar Haldorai and [2]Shrinand Anandakumar
[1]Department of Computer Science and Engineering, Sri Eshwar College of Engineering, India.
[2]The PSBB Millennium School, Coimbatore, India.
[1]anandakumar.psgtech@gmail.com, [2]shrinand.psbb@gmail.com

Article Info
Journal of Computing and Natural Science (http://anapub.co.ke/journals/jcns/jcns.html)
Doi: https://doi.org/10.53759/181X/JCNS202202011
Received 25 December 2021; Revised form 20 March 2022; Accepted 25 April 2022.
Available online 05 July 2022.

Abstract – The term " Edge Artificial Intelligence (Edge AI)" refers to the part of a network where data is analysed and aggregated. Dispersed networks, such as those found in the Internet of Things (IoT), have enormous ramifications when it comes to "Edge AI," or "intelligence at the edge". Smartphone applications like real-time traffic data and facial recognition data, including semi-autonomous smart devices and automobiles are integrated in this class. Edge AI products include wearable health monitors, security cameras, drones, robots, smart speakers and video games. Edge AI was established due to the marriage of Artificial Intelligence with cutting Edge Computing (EC) systems. Edge Intelligence (EI) is a terminology utilized to define the model learning or the inference processes, which happen at the system edge by employing available computational resources and data from the edge nodes to the end devices under cloud computing paradigm. This paper provides a light on "Edge AI" and the elements that contribute to it. In this paper, Edge AI's motivation, definition, applications, and long-term prospects are examined.

Keywords – Artificial Intelligence (AI), Edge Artificial Intelligence (Edge AI), Edge Computing (EC), Internet of Things (IoT), Machine Learning (ML).

I. INTRODUCTION

Over the past few decades, the application of Artificial Intelligence (AI) has significantly increased around the globe. Due to the expansion of commercialized activities in business, the cloud computing paradigm has become a pivotal element of AI's progress. More businesses are realising the need of making their technology available on smartphones so that they may better meet the needs of their customers. As a result, growth in the Edge Computing (EC) [1] business is projected in the next years. Edge Artificial Intelligence (Edge AI) analyses data generated by a local hardware device using Machine Learning (ML) algorithms. Real-time millisecond processing and decisions do not need an Internet connection. As a consequence, the communication costs of the cloud model are drastically lowered. Data and processing are two areas where Edge AI focuses its efforts, moving datasets and processing nearer to an interaction point with users. Examples of this technology in action include Google's Homepod, Amazon's Alexa, and Apple's Homepod, all of which employ ML to learn and recall local words and phrases. Artificial Intelligence (AI) analyses the user's inquiry through an Edge network, which translates the user's speech into text. With no edge networks, the time of respond would take longer; with edge, the time will be 400 milliseconds longer than normal.

Whenever it comes to Edge AI or cutting-edge AI [2], the present study is still in its infancy. In order to demonstrate the benefits and applications of Edge AI, Microsoft created a mobile voice command recognition prototype in 2009. Although research on Edge AI is still at an early stage, there is still no precise definition of the term. "Edge AI" is currently being used to describe the practise of running AI algorithms directly on an end device utilising data produced on the device itself (sensor data or signals). Despite the fact that Edge AI is now the most often used method, it is important to remember that this definition severely limits the range of possible solutions (e.g., using high-end AI processors). High-end CPUs are required to run DNN models, for example, locally due of their computationally demanding nature. Additionally, these rigorous criteria are incompatible with existing legacy end devices, which have a limited computing capacity, pushing up the cost of Edge AI further.

As we explain in this section, Edge AI need not to be confined to operating the models of AI just on the edge devices or servers. According to a dozen recent research, running Deep Neural Network (DNN) models [3] with edge-cloud synergy decreases end-to-end delay and consumption of energy in contrast to the operating them on a local level. There are various purposes as to why we have a belief on a collaborative hierarchy being integrated into the establishing of cutting-edge information gathering technology systems. Training and inference are assumed to take place in power cloud datacenters, and this assumption is widely held. Since training consumes much more resources than inference, this is a reasonable assumption to base our design decisions on.

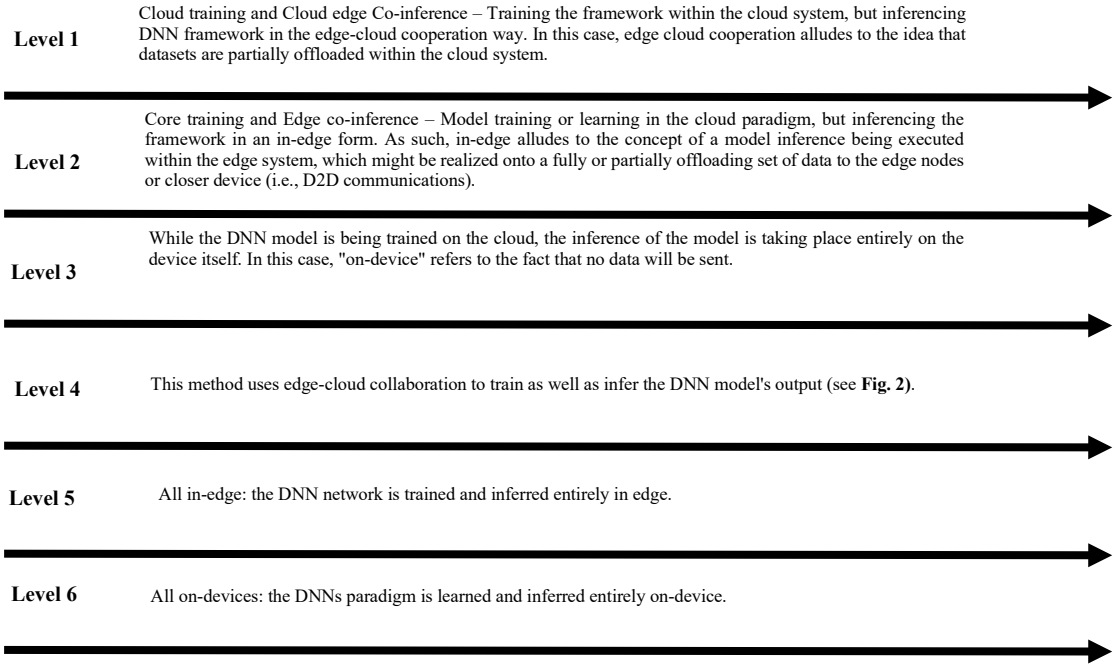

Fig 1: Six levels of Edge AI

This would need sending enormous amounts of training data from the edges or the AI devices to the cloud computing system [4] to avoid any prohibitive communications privacy or cost issues. Based on the forward-looing view point, we have a belief that Edge AI has to be a paradigm, which completely makes use of the prevalent resources and datasets throughout the hierarchies to the edge node, cloud datacenters and the end devices to potentially enhance the general performance of inference and learning the DNNs model. There may be no need for data offloading between the cloud, edge, and devices for the DNN model, since it may be trained or inferred entirely at the edge. Data dumping volume and route length are used to determine the six levels of Edge AI shown in **Fig. 1**. The following is a breakdown of the many levels of Edge AI.

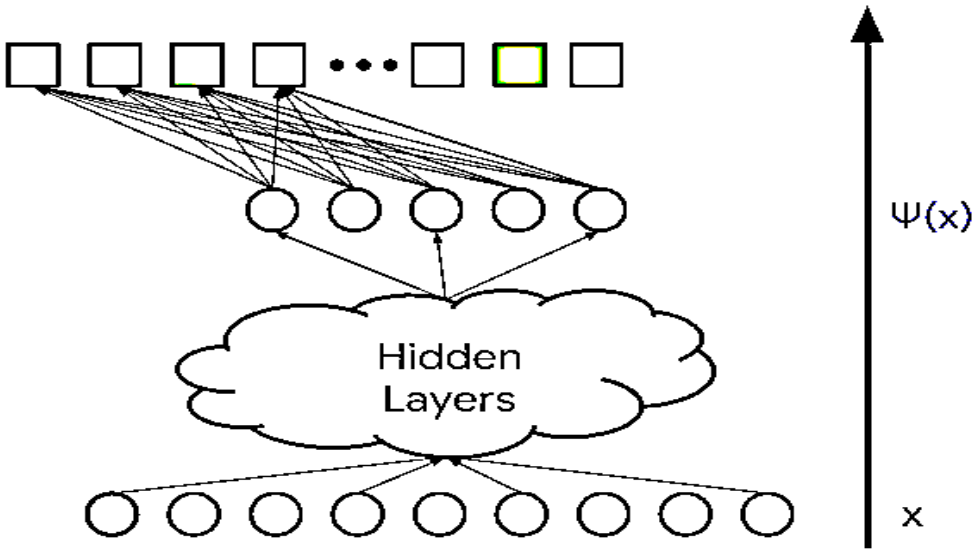

Fig 2. DNN output in the hidden layer $\psi(x)$

Since Edge AI improves, it becomes more obvious that the volume and route length for data offloading decrease. Data offloading latency is lowered, data privacy is increased, and WAN bandwidth costs are reduced as a result of this. However, the price of this enhancement is greater computational latency and energy usage. Application-specific Edge AI should be developed by taking into account factors e.g., WAN bandwidth costs, privacy, energy efficacy and latency whenever determining the appropriate amount of Edge AI. This paradox illustrates the fact that there is no "one-size-fits-all" approach to problem solving. The ultimate goal and future of computers, in our opinion, is level-6 Edge AI. For different levels of Edge AI, commercial solutions and enabling methods will be described in the following chapters. Edge AI was created as a

result of the marriage of cutting-edge computers with AI. Using data and computing resources from the cloud datacenters, to edge nodes and end devices to train or infer models at the network edge is what is meant by "Edge AI".

Edge AI's motivation, definition, applications, and long-term prospects are all examined in this research. In that regard, the content is arranged as follows: Section II focusses on the motivation of the research while Section III presents a critical definition of Edge AI. Section IV reviews the advantages of Edge AI in different real-life contexts. Section V focusses deeply on the fundamental application of Edge AI. Section VI projects the future of Edge AI in a web-linked ecosystem while Section VII presents concluding remarks on the paper.

II. MOTIVATION BEHND THE STUDY

This article discusses the benefits of Edge AI and how to put it to use. When it comes to applications, edge AI technology has no boundaries. Because of the Covid-19 conundrum, organisations have turned to AI to offer real-time data solutions. As an example, AI is being used in the monitoring, assessment, and treatment of patients, for example. The developer community's creativity and imagination are the only real limitations of Edge technology. Because of this, a number of collaborative projects are already underway to educate Science, Technology, Engineering and Mathematics (STEM) professionals and students about this new technology. A joint effort between Intel and Udacity is developing the Intel Edge AI for IoT Professionals Nanodegree initiative, which teaches computer vision and deep learning. This category includes software developers, ML technologists, data experts, and professionals engaged in the creation of cloud-centered AI devices and applications. Due to the programme, it is projected that EI applications would become more user-centric.

III. DEFINITION OF EDGE AI

The term "Edge Computing" refers to any programme that reduces latency closer to the user's request. By "Edge Computing," Li and Hong [5] refers to computation that occurs outside the cloud, at the network's edges, and in applications that need real-time processing of data. They explained this in his keynote talk to IEEE DAC 2014 and in an invited session at MIT's MTL Seminar in 2015. EC, in contrast to cloud computing, makes use of data that is constantly being generated by sensors or users.

Edge AI is the most popular term used to describe fog computing. When it comes to the EC concept, servers located in the "final mile" of a network are regarded to be a part of it. An appropriate reference is required. According to [6], anything that is not a normal data centre may be an 'edge. As an edge node, gamelets broadcast games to clients that are one or two hops distant. In order to meet the response time requirements for real-time gaming, edge nodes are often one or two hops removed from the mobile client. In the context of EC, virtualization technologies may make it easier to set up and manage many different types of applications on edge servers.

When data is analysed and combined at the network's outer edge, it's called an "Edge AI." Using the term "Edge AI," or "intelligence at the edge," has significant ramifications when it comes to scattered networks like the IoT. System nodes placed outside the system's core may now execute functions that were previously only available to the system's core. Traditional data storage and retrieval from IoT-connected devices and sensors in a central data warehouse or repository has a number of disadvantages. It is possible that the system will be more vulnerable and inefficient if the data is not secured. These nodes may intelligently analyse the data as it moves through the edge network and into the data warehouse in a smart edge architecture. Because of this, data-handling systems may become more responsive and safer. As a result of these and other factors, a number of cloud providers and other enterprises with experience in IoT structure and nature advocate the use of an Edge AI.

IV. BENEFITS OF EDGE AI

This includes smartphone apps such as feature extraction and real-time traffic dataset; and semi-automated vehicles or smart gadgets. Gaming consoles, smart speakers, robotics, drones and surveillance cameras are among the Edge AI-enabled gadgets. Here are a few additional applications for Edge AI that we hope to see in the future:

- It will aid in the identification of security cameras by providing information. Traditional video security cameras capture long-exposure photos that may be stored and retrieved as needed. It's a different story with Edge AI, which uses real-time algorithmic processes to identify and analyse suspicious actions in real time, making it less costly and efficient for monitoring.
- Increased real-time processing of data and photos will allow autonomous cars to recognize traffic signs, people and other vehicles as well as roadways. This will improve transportation security.
- For example, it might be used to analyze images and videos and create reactions to audio-visual stimuli or real-time locale or scene identification in smartphones.
- Minimize the costs and enhance safety with IIoT (Industrial IoT). Edge AI will survey machines for suspected flaws or mistakes, while ML will assemble real-time data from the whole manufacturing chain.
- It will be utilized in the field of emergency medicine to analyze photos.
- The development of 5G technology networks will lead to faster and more reliable mobile data transfer, which will make Edge AI more effective. For example, Red Hat and IBM have collaborated to establish 5G-oriented edge technologies; hence making it easy for firms to effectively manage activities across larger numbers of devices from multiple suppliers, providing communications firms the agility it required to promptly provide the required services to consumers.

Since EC and AI have many commonalities, it seems sense that the two will come together. Artificial Intelligence (AI) and EC, on the other hand, try to mimic the behaviors of machines and devices by learning from data established by edge servers and devices. Additionally, to obtaining the rewards of EC, the following advantages accrue when AI is pushed to the limit.

Data created at the periphery of the network requires AI to unlock its full potential. End devices of many kinds, from security surveillance systems, wearables sensors and smart sensors, to IoDs (Internet of Drones), have been interconnected with the web in recent times. It is shown in **Fig. 3** that the Internet of Things (IoT) setup is utilised for connecting devices and sensors to the web directly. In this scenario, raw data is sent to backend servers for processing by ML algorithms.

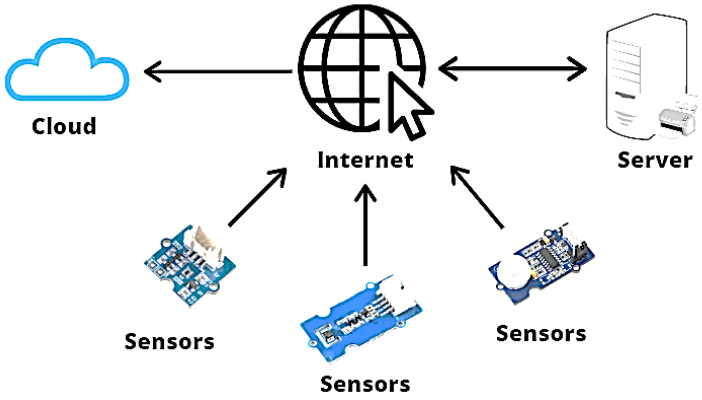

Fig 3. IoT configuration

In **Fig. 3**, IoT configuration is used for the connection of devices or sensors directly to the web. In this case, raw data is provided to backend servers whereby ML algorithms are operated.

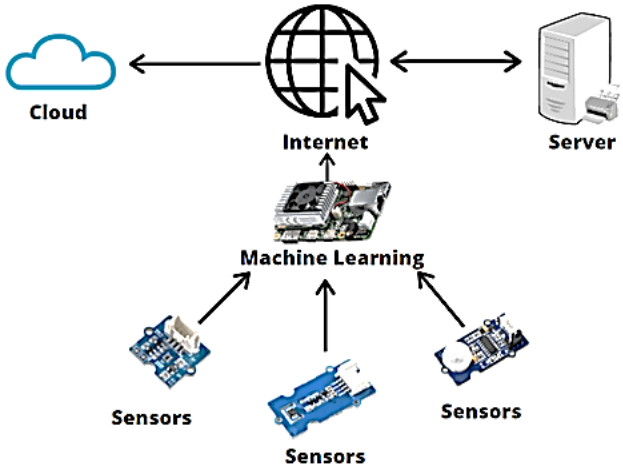

Fig 4. Edge AI configuration of machine learning

Fig. 4 shows configuration of Edge AI run locally by ML algorithms on the embedded systems or hardware devices in contrast to the servers. Based on the AI application or the category of the device, there are different hardware categories for the performance and activity of edge AI processing e.g., SoC, FPGAs, ASICs, GPUs, and CPUs accelerators.

A considerable amount of data (such as video, image, and audio) is continually detected at the device end as a consequence of the proliferation of these different gadgets. The potential of AI to swiftly evaluate and derive insights from such vast amounts of data will make it a functional need in this environment. With deep learning, which is one of the most widely used AI approaches, the edge device can automatically discover patterns and detect abnormalities based on the data it collects from the surrounding environment. Real-time predictive decision-making based on sensed data (e.g., planning for public transit and traffic management) is therefore able to respond more quickly and effectively to rapidly changing surroundings. Deep learning algorithms outperform classical intelligence methods that rely on the tracking of numerical limits to be fulfilled. PTC ThingWorx, Amazon AWS IoT, IBM Watson IoT, IBM Watson IoT, and the Microsoft Azure IoT, are the leading IIoT systems, which are embedded with predictive AI capacities. In contrast with today's 10%, Liu et al. [7] project that by the end of 2022, more than 85% of firms' IoTs initiatives will integrate an AI element.

EC, on the other hand, allows AI to be enriched with large amounts of data and application scenarios. Deep learning's current rise has been attributed to four factors: algorithm, hardware, data, and application contexts. Data and application

scenarios have received much less attention than algorithmic and hardware factors in deep learning research. The most popular method for improving the performance of deep learning algorithms is to add extra layers of neurons to the DNN. We need to learn additional DNN parameters, and the amount of data needed for training increases as a result of this. We can see that the average delayed time between significant methods and accompanying advancements over the last 30 years has been approximately eighteen years, wherein similar advancements between vital datasets have taken less than 3 years on average.

This is a clear demonstration of the relevance of data in the advancement of AI. Following the recognition of the value of data, the following question is, "Where do we get our data from?" Mega-scale datacenters have traditionally been the primary source of data generation and storage. However, with the fast expansion of the IoT, this tendency is gradually being reversed. Chen et al. [8] predicted that by the end of 2022, all people, machines, and things will create roughly 850 ZB, up from 220 ZB generated in 2016.

Smart finance, cancer diagnosis, and drug development are just a few of the new frontiers that AI is helping to unlock. The objective of "developing AI for each and every business at any moment in time" has been expressed by major IT firms, with a broad variety of applications and possibilities. This requires bringing AI up close to the people, information, and edge devices. When it comes to accomplishing this goal, EC is definitely better than cloud technology. Edge workstations, as contrasted to cloud data centres, are located nearer to end users, data sources, and other computing equipment than cloud data centres. As a result, EC is cheaper and simpler to adopt than cloud computing systems. Can be used to get around obstacles

AI technologies, on the other hand, may help popularise EC even more. There has always been an issue to worry about in the cloud computing realm whereby high-demand services edge technology might take to the next level, which cloud computing could not. To dispel any confusion, Microsoft's research team, which co-invented the cloudlet concept, has been investigating since 2009 what sorts of applications should be transferred from the clouds to the edge, including voice control identification, AR/VR, immersive cloud gameplay, and real-time video monitoring. A much more compelling use case for cutting-edge computers would be real-time video analytics. Using machine vision, real-time video intelligence continually extracts high-definition movies from security camera feeds and analyses them, necessitating a large amount of computing, high bandwidth, confidentiality, and reduced latency. Evidently, EC is the only conceivable option that can achieve these stringent criteria. A dozen cognitive support applications, such as machine vision, voice recognition, and computational linguistics, have been the primary focus of CMU experts in their efforts to promote the cloudlet idea. The rise of EC may be traced back to the rise of AI applications, which have played a significant influence in its popularity. Most mobile and IoT-related AI solutions have this in common since they are computationally and energy expensive, privacy-sensitive, or otherwise affected by latency or other delays. This makes them a good fit for EC.

Edge AI has gotten a lot of attention lately because of its advantages and requirement in operating AI applications at the edge. Cloud-edge AI systems were proposed as a significant research topic in December 2017 by the UC Berkeley (University of California, Berkeley) in the paper: “A Berkeley View on Systems Constraints for AI”. 'Edge AI' made its debut in Gartner's Hype Cycle in August of 2021 [9]. In the next 5 to 10 years, Gartner predicts that edge AI will hit a plateau of efficiency, which is still in the development trigger stage. Many pilot projects for cutting-edge AI have been carried out in the sector as well. The conventional cloud service providers, e.g., Google, Microsoft, and Amazon have built service frameworks to extend the knowledge of Cloud Servers to the edge, allowing network nodes to conduct ML inferences locally using pre-trained models. Google Edge TPU, Huawei Ascend 910, Intel Nervana NNP, and Ascend 310 are four examples of high-end AI processors designed for executing ML models currently available on the marketplace.

V. APPLICATION OF EDGE AI

AI and Retail: Shopping Experience

It is critical for small and big merchants alike to provide their customers with an enjoyable buying experience, since it has a significant impact on consumer loyalty. With the help of AI, companies can keep consumers happy and keep them coming back for more transactions. Another one of the many Edge AI implementations used to assist workers in their day-to-day operational processes and to improve the client experience is using AI to ascertain when product lines being sold need newer products and when product lines have to be resupplied (mostly appropriate for perishable commodities). Smart checkout structures utilise ML (a branch of ML that assists computer systems in determining the characteristics of objects, classifying them, and drawing inferences from digital images acquired from video content and webcams) to verify that the barcodes on the objects being scanned are those that belong to the items being identified. Smart surveillance analytics is also used by retailers to learn more about their consumers' preferences and better plan their shop layouts. Customers' buying habits are also analysed by merchants in order to enhance their consumer shopping experiences by evaluating data about transactions and abstracted information from videos.

AI in Smart Industries: Business Experience

Fast Moving Consumer Goods (FMCG) precision manufacturing companies must constantly guarantee the accuracy and safety of their products. The manufacturing plant will be safer and more effective as a consequence of the use of AI and EC in firm operations. In-plant inspections may be carried out using Edge AI applications, such as those created by BMW and Procter & Gamble. In order to help Procter & Gamble employees, the company uses an inspection camera. Video footage is

analysed to ensure that no defective products leave the factory. Quality control and safety initiatives are bolstered as a consequence. For instance, Edge AI can detect quality assurance, verification, and deviation issues that the human visual system could overlook. Industrial quality control may be effectively monitored using computer vision. Other automakers are also using EC and AI on the assembly line for real-time views and insights. This has resulted in a more efficient production process.

Powering Smart Hospitals: Medical Service Experience

A constant need exists for healthcare experts to require technological assistance from time to time. Using AI and cutting-edge computers in the medical profession will improve patient care and operational effectiveness. Data security is a major concern for smart hospitals, and edge AI applications can assist with that. It is possible to undertake high accuracy thermal scanning, inventory control, remote patients surveillance and even disease prediction with Edge AI in the medical industry. Smart healthcare facilities may become a reality with the help of Edge AI apps.

The COVID-19 pandemic and the emergence of 5G networks, as well as the growing interest in telehealth and other connected solutions, have prompted more people to consider the notion of a smart hospital in recent months. Medical professionals who have access to cutting-edge technology will be able to rapidly and readily obtain patient health measurements, test results and other information in a smart hospital. A renowned data and analytics business called GlobalData [10] argues it will be necessary to coordinate technologies from a number of areas to make this a reality.

The rising use of robotics in hospitals is an excellent illustration of the necessity for smart hospitals to have interconnectivity across many technologies. Medication and medical equipment delivery robots, assistive robots for doctors and nurses to transfer patients or execute operations, and surgical robots are some examples. In order to keep the hospital running well, all of the robots in the facility must be able to be supervised and controlled at all times. In hospitals, fewer people would be needed to supervise the network of robots since computer software would take care of it first.

The smart hospital's network infrastructure will be its most important component. Intrahospital networks will be required to swiftly exchange, transport, and distribute data between equipment and departments in these hospitals. There will be a wide range of medical instruments in the hospital, including on-patient sensors and magnetic resonance imaging scanners, that will require ongoing connectivity. However, 5G might eliminate these issues while allowing for faster interhospital links. These may then be utilised for patient or physician discussions, enhancing treatment efficiency.

It's currently possible to build smart hospitals, but it will need a big investment from all parties. Manufacturers of medical equipment must make sure that new products are both functional and safe from cyber-attacks, while also being simple to use for the patient. Hospitals will need to focus on areas that provide the greatest value, while ensuring that their fundamental IT infrastructure is safe and adequate to meet the needs of their patients. To take advantage of 5G and the smart hospital, telecoms will have to make significant investments in network improvements.

Telecommunications: Communication Experience

For telecommunications companies, Edge AI Applications can provide different and innovative experiences in the application of 5G for quality assurance checking in factories equipped with sensors and cameras, in the manner that software-defined systems are utilised to computerise self-checkout processes in stores, and for customer perception powered by AI. Network operators will benefit greatly from the combination of these factors.

Drones:* Visual *Analysis

In addition to traffic control, engineering, and mapping, drones equipped with edge AI may be utilised for a variety of other purposes. Image classification, document classification, and object identification and tracking are all common features in drones today. Drones equipped with AI are able to identify and locate objects by replicating human visual search patterns. As a result of using edge AI technologies in drones, drone data may be successfully examined. In addition to facial object recognition and monitoring in real-time, it helps with predictive management and logo identification, as shown in **Table 1**.

Fire Fighting Robots

Similar to how drones are used to inspect dangerous regions that are too dangerous or difficult for humans, robots may be utilised to put out fires. As they go around a structure, AI-enabled robots may inspect and map it for structural damage while saving lives and extinguishing fires. The Howe and Howe Innovations Thermite Robot is an illustration of a firefighting robot, which was initially built for the US Army and includes cameras and a water line that can pump 500 gallons per minute linked to it. With Edge AI installed, it can access risky areas without the need of a remote control and can recognise harmful situations that humans are unable to encounter and take action to relieve them.

Autonomous Vehicles

Autonomous vehicles (also known as self-driving automobiles) are a great example of cutting-edge AI. Data processing may be done on the same hardware as the device itself with Edge AI-enabled devices [11]. Preventing fatal accidents requires the processing capability included inside the same hardware. If the process is too sluggish, it might result in severe consequences. Data from this Edge AI application is collected and analysed in real-time in order to assist drivers and other

road users avoid accidents and avoid potential hazards. Automation is not a new phenomenon in commercial aviation, since sensors are constantly being monitored and analysed in order to ensure the safety of the aircraft.

Table 1. Benefits of drones for visual analysis

Significance	Definition
Forewarned repairs	It is possible to do preventative maintenance on old and deteriorating infrastructure in order to avoid collapse and total erosion. Structures, bridges, and roadways, for example, deteriorate and lose strength with time, posing a danger to many people if they are not adequately maintained in a timely manner. There are several ways in which drones may assist in ensuring that the correct structures are repaired quickly in order to prevent further deterioration.
Detection of logos	There are a variety of uses for drones in marketing, including gathering data on advertising campaigns or logo placements to better understand how such efforts are being received by target audiences.
Real-time object identification and tracking	Drones may be used to keep an eye on congestion and detect and track moving objects in real time. Tracking traffic violators and fugitives ensures security and safety.
Face recognition	Drones equipped with AI may be used to monitor pedestrian movement, particularly in high-risk areas with a large volume of traffic. A variety of applications, including law enforcement, call for its use.
Cartography and Mapping:	Using drones for cartography and mapping is advantageous since they are able to access locations that are inaccessible to people. An AI-powered drone can achieve what a team of professionals would have to do in a fraction of the time and at a fraction of the expense. To create 3D pollution mapping, they can collect data on things like noise, air, or radioactive pollution, and they can scan things like bridges, structures, and roads to see whether damage has occurred.

Industrial IoT (IIoT)

The Industrial IoT relies heavily on the digitalization of operational operations and production in order to increase productivity. Using Edge AI, industrial IoT devices can do visual examination and manage robots more quickly and at reduced prices.

Energy

The usage of smart grids is one example of edge AI uses in the energy industry, which may be seen. These smart grids can use renewable energy, monitor usage and decentralise energy production because of EC driven AI. In order for smart grids to work, data must be sent directly between devices, hence it is preferable not to use the cloud for this [17-19].

Safe, Smart Road Networks: Transport Experience

Roads are becoming smarter in an effort to make them safer for drivers in the future. In addition to establishing smart roads, some tech individuals and organizations are hoping to create entirely smart cities where roads, cars, and buildings all communicate with one another. Edge AI techniques are frequently used to offer cameras that study traffic in real time to detect obstructions, irresponsible driving, disasters, and hold-ups. Edge AI solutions are also being used by businesses to improve traffic flow and make roadways safer for all road users, including walkers, cars, motorcyclists, and others. There are also edge AI applications in surveillance cameras, which do not need to send raw video signals intended for computation to the cloud. It's possible for Edge AI-powered cameras to analyse photos on the spot and in real time. Using this technology, it is possible to monitor and track things in order to take rapid action and to prevent fatalities from occurring [20-22].

VI. THE FUTURE OF EDGE AI

We conduct a complete review of the literature on cutting-Edge AI and provide our findings. Because it allows for the development of AI in the last mile and enables consumers to benefit from high-efficiency intelligent services without having to rely on central cloud servers, Edge AI is an apparent plus. Reiterating that there are still unresolved issues in achieving Edge AI is a good idea. Identifying and analysing these problems is essential, as is looking for new theoretical and technological answers. This Some of the most pressing issues in Edge AI are discussed here, along with some potential solutions. Problems with data scarcity at the edge, inconsistency on edge devices, poor flexibility of statically trained models and incentive mechanisms are only a few of the difficulties faced [23-25].

Data scarcity at edge

Because of the need on high-quality training data for most ML algorithms — particularly supervised ML — high performance is a pre-requisite. HAR and speech-recognition applications, for example, are examples of Edge AI settings where the gathered data is sparse and unlabeled. In contrast to typical cloud-based intelligent services, which collect all of the training data into a single database, edge devices develop models from data they collect or data they produce themselves. Good picture features, for example, are missing from these datasets. Assuming that the training examples are of high quality,

most current efforts disregard this problem. In addition, the training data is often unlabeled. To address the issue of unlabeled training examples, several studies advocate the use of active learning.

In cases when there are just a few examples and categories, this strategy may be employed. The decentralised nature of data may be used in a federated learning strategy to successfully tackle the issue. Federated learning, on the other hand, is best suited for group training rather than the one-on-one instruction required for personalised models.

The following is a list of potential remedies to this issue.

- Adopt models that can be trained with a limited amount of data. In general, a ML method that is easier to learn from tiny datasets would perform better. In certain cases, a basic model, such as Naive Bayes, linear model, and decision tree, is sufficient to cope with the issue since they are simply attempting to learn less. As a result, when confronted with real-world issues, it's important to choose the right model.
- Incremental techniques of learning. In order to integrate their fresh data, edge devices might retrain a regularly used pre-trained model in an unprecedented manner. Fewer training sessions are needed to create a bespoke model in this approach.
- Methods based on transfer learning, such as few-shot learning.
- The cold-start issue is often avoided, and the quantity of training data needed is reduced, thanks to the use of transfer learning, which applies what has been learnt from one model to another. Transfer learning may be an option if the target training data is few and the source and target domains share a few characteristics.
- Methods based on data augmentation. To enhance the effectiveness of a specific model, data augmentation is used throughout the training process. For instance, flipping, rotating, scaling, translating, cropping, and the like may increase the number of photos while maintaining the semantic meaning of the labels. The network's performance on new data will be improved because of the training on enhanced data, which will make it more resistant to deformations.

Data consistency on edge devices

Edge AI applications, such as voice recognition, event detection, and face recognition, typically use sensors at the network's edge to collect data. If the collected data is not consistent, it could have an impact on the results [12]. Since sensors can be found in so many different places, this problem is made worse. Environmental noise (e.g., wind or rain) and its conditions can have an impact on the sensor data collected (e.g., library or street). Due to the heterogeneity of sensors, data collected by sensors may also be subject to variation (e.g., hardware and software). In terms of sensitivity, sampling rate, and sensing efficiency, sensors can differ greatly. Sensor data collected from the same source may be interpreted differently by different sensors.

Training parameters like feature variables will change as the data changes because they are dependent on the training data. Despite the wide ranges of variance, existing sensing solutions are still hampered. As long as the model is taught centrally, this problem can be easily solved. A large training set is centrally located to ensure that the invariant properties of the variants can be learned. While this situation isn't covered by Edge AI, it's a common one. It is imperative that future attempts to fix this issue take into account the detrimental effect that variance has on model accuracy. We may explore data augmentation and reinforcement learning in this regard. Data augmentation may be used to improve the model's capacity to handle noise. As an example, a variety of background noises may be used in speech recognition applications for mobile devices in order to avoid the varying effects of the environment. As an analogy, sensor hardware noise might be used to solve the problem of inconsistent findings. Because of the usage of enhanced data, the algorithms are better equipped to handle these changes.

The representation of data has a significant impact on model performance. There are many different approaches to learning the representation of data in order to extract more useful characteristics when building models. If we could translate between two sensors that use the same data source, we'd see a massive boost in the model's performance. As a result of this, representation learning has the potential to reduce the detrimental consequences of inconsistent data. Possible future developments in this area include more efficient pipelines and data processing.

Bad adaptability of statically trained model

First, the system is trained on a centralized server, and then distributed to edge devices in most AI applications. Once the training phase is complete, the model will not be retrained. Users will have an unpleasant user experience because of the poor performance of statically trained systems when confronted with unexpected data or tasks. Decentralized learning, on the other hand, simply takes into account data from a single location. As a result, these models may only become specialists in their limited field. The level of service suffers as the region served expands. Lifelong deep learning and knowledge exchange are two potential answers to this challenge.

Continuous learning and self-improvement are possible with the LML paradigm, which is a more sophisticated approach to learning. In the future, machines will be able to learn new things on their own rather than being taught by humans. LML differs from meta learning, which allows computers to learn new models autonomously. To cope with new data and changing contexts, edge devices might employ LML and a sequence of learnt tasks. Remember that the LML is not meant to be used by low-end devices, which means that they will need to have a high level of processing power. As a result, if LML is used, model design, model decompression, and offloading mechanisms should all be taken into account.

Knowledge sharing facilitates the exchange of information between several edge servers. One way an edge server may get help when it receives a job it can't do is by sending knowledge requests to other network edge. Because the knowledge is distributed among several network edge, the server that has the necessary information answers to the request and completes the job for the user. In a knowledge sharing model like this, a technique for assessing and querying one's own expertise is necessary.

Privacy and security issues

Diverse smart objects and edge servers must work together to supply computational power in order to realise Edge AI. There's a chance that the publicly cached data and computing jobs (either learning or inferences tasks) will be transmitted to unfamiliar machines in this operation for further processing. As a result of the data's potential to contain personally identifiable information, such as photos and tokens, it raises the possibility of data breaches and attacks by malicious users. Without encryption, unscrupulous people have easy access to sensitive data. To hide personal data and compress the data conveyed, some initiatives posit to perform some preparatory processing locally. However, the processed data may still be used to get private information. Viruses inserted into computer processes may also be used by malevolent people to attack and take control of computing devices. Users' privacy and security are at risk due to the absence of appropriate privacy and security policies or methods.

Another option is to use a credit system. Like the commercial bank system, which verifies each user and verifications their credit information, this is a similar system. Creditors with bad records would be removed from the database. Because of this, all computers that supply processing power are trustworthy and safe for all users. Several works e.g., by Han, Pan, and Li [13] have previously employed encryption to secure their subjects' private information as a privacy safeguard. However, the encrypted data must be decrypted before training or inference activities can be carried out, which necessitates an additional amount of processing. This issue could be addressed in the future by increasing the focus on homomorphic encryption. Direct computations on encoded ciphertexts may create encrypted outputs using homomorphic encoding. There is no difference in the outcome of decryption and calculation on unencrypted data 42. Homomorphic encoding, on the other hand, makes it possible to directly perform training or inference tasks on encrypted files.

Incentive mechanism

Edge AI relies heavily on data gathering and model training/inference. It is difficult to guarantee the accuracy and usefulness of the obtained data while it is being used for data gathering. All the resources and time needed to perceive and gather data are consumed by data collectors. Preprocessing data cleaning, extraction of features, and encryption need additional resources that can't be assumed to be shared by all data collectors. Participants in a collaborative model training/inference must work together unselfishly on a particular task. AI architectures such as [14]'s have one ‘master’ and many ‘slaves’, for example. Pipelines allow workers to recognise items in a certain mobile visual environment and offer masters with training examples. This kind of architecture works best in private settings, like a person's house, where all the equipment is compelled to work together to build a superior intelligent modeling for their proprietor.

However, in situations when the master initiates a task and assigns subtasks to unknown players, this approach would fail. Typically, in smart environments where all devices are not owned by a single master, an additional incentive issue arises. To encourage data collection and task completion, participants must be rewarded. Efforts in the future should examine the use of reasonable incentive systems. Data gathering, data management, and data processing all require different amounts of resources from participants. Everyone who takes part is hoping to receive the biggest award imaginable. The operator, on the other hand, is looking for the highest predictive performance at the lowest feasible cost. The most difficult part of building the best incentive system is figuring out how to measure the workloads of various missions to match the related rewards. These problems might be the focus of future initiatives.

Predictive AI

Predictive analytics powered by AI [15] is helping many firms better forecast customer behaviour and take preventative steps. Predictive AI built on Apache Hivemall, for example, may be used by a customer data platform (CDP) to assess consumers based on factors like churn affinity or upselling potential. Marketers may then target specific consumers based on these rankings. Businesses in various industries might use this as a source of inspiration. For example, a top game developer was able to accurately forecast, using Treasure Data's CDP and ML, which sorts of in-game prizes would keep players interested. Pretend there's a retail business out there attempting to hold onto its most loyal clients. AI-based predictive scoring may be used to identify and alert customers who are most likely to discontinue their purchases. Predictive AI may be able to identify customers who are about to leave a purchase and then learn over time which offers or interactions tempt them back.

With the help of artificial intelligence, it is possible to determine which characteristics and patterns of behaviour suggest that a consumer is on the verge of making a purchase. The system may issue an invitation to an in-store demo when it recognises a consumer who meets a certain profile and therefore assist enhance sales. Predictive AI in the back office will have an impact on customer service. Inventory management systems will enhance their ability to search for predicted trends and select when and where to distribute products, similar to what Amazon is currently doing with AI-based approaches like anticipatory delivery. As a consequence of this new trend, customers will have greater in-store options and reduced wait

times. Retailers could see more happy customers if they can provide faster, simpler, and more convenient in-store pick-up choices to their customers.

In the near future, new automated ML tools may make predictive AI available to both small and large organisations, democratising the technology. Northstar, an interactive prediction tool, was developed by MIT and Brown University researchers [16]. On any touchscreen device, Northstar's drag-and-drop graphical interface lets users to input datasets and build predictive ML models. With predictive AI, any organisation, no matter how large or small, may benefit without having to hire in-house data scientists. Using Northstar, a small business owner may estimate sales based on historical data, and then choose which items they want to keep on hand.

VII. CONCLUSION

With the application of distributed Edge Computing (EC), data storage and processing are nearer to the source as opposed to traditional cloud computing. Reaction times are expected to improve, but bandwidth is expected to decrease. In this case, there are two ways to look at it: "Misconceptions about Internet of Things (IoT) and EC abound. Using EC, the IoT is an example of how this type of distributed computing can be used." It's an architectural term, not a specific technology. To deliver web and video content from servers close to users, content dispersed networks were created in the late 1990s. This is where the origins of EC can be traced back to. It wasn't until the early 2000s that services such as shopping carts and dealer locators with real-time aggregates of data and ad insertion machines were developed for commercial use. This paper has focussed on Edge Artificial Intelligence (Edge AI) motivation, definition, applications, benefits and the future. As an example, this research demonstrates how edge AI can improve customer experiences, eliminate human risk, supplement human healthcare efforts, and make roads safer. In a wide range of industries, businesses are relying on edge AI applications to improve operational efficiency and real-time monitoring. Edge AI protects data, maintains privacy, eliminates latency and bandwidth issues, and lowers hardware costs. Using Edge AI applications in your company requires a willingness to embrace new technology and an understanding of business practises. With AI, you may make use of a wide range of sensors, including those found in drones, robots, inspection cameras, and many more.

References

[1]. R. Meneguette, R. De Grande, J. Ueyama, G. P. R. Filho, and E. Madeira, "Vehicular Edge Computing: Architecture, resource management, security, and challenges," ACM Comput. Surv., vol. 55, no. 1, pp. 1–46, 2023.

[2]. G. Fortino, M. Zhou, M. M. Hassan, M. Pathan, and S. Karnouskos, "Pushing Artificial Intelligence to the Edge: Emerging trends, issues and challenges," Eng. Appl. Artif. Intell., vol. 103, no. 104298, p. 104298, 2021.

[3]. A. B. P. Samson, S. R. A. Chandra, and M. Manikant, "A deep neural network approach for the prediction of protein subcellular localization," Neural Netw. World, vol. 31, no. 1, pp. 29–45, 2021.

[4]. X. Tang, Y. Liu, Z. Zeng, and B. Veeravalli, "Service cost effective and reliability aware job scheduling algorithm on cloud computing systems," IEEE trans. cloud comput., pp. 1–1, 2022.

[5]. Y. Li and Y. Hong, "Prediction of football match results based on edge computing and machine learning technology," Int. j. mob. comput. multimed. commun., vol. 13, no. 2, pp. 1–10, 2022.

[6]. V. D. A. Kumar, A. Kumar, R. S. Batth, M. Rashid, S. K. Gupta, and M. Raghuraman, "Efficient data transfer in edge envisioned environment using artificial intelligence based edge node algorithm," Trans. emerg. telecommun. technol., vol. 32, no. 6, 2021.

[7]. S. Liu, C. Guo, F. Al-Turjman, K. Muhammad, and V. H. C. de Albuquerque, "Reliability of response region: A novel mechanism in visual tracking by edge computing for IIoT environments," Mech. Syst. Signal Process., vol. 138, no. 106537, p. 106537, 2020.

[8]. Y. Chen, W. Tong, D. Feng, and Z. Wang, "Cora: Data correlations-based storage policies for cloud object storage," Future Gener. Comput. Syst., vol. 129, pp. 331–346, 2022.

[9]. "Presidential working group on artificial intelligence," Ucop.edu. [Online]. Available: https://www.ucop.edu/ethics-compliance-audit-services/compliance/presidential-working-group-on-artificial-intelligence.html. [Accessed: 08-Mar-2022].

[10]. "ShieldSquare captcha," Globaldata.com. [Online]. Available: https://www.globaldata.com/. [Accessed: 08-Mar-2022].

[11]. M. Merenda, C. Porcaro, and D. Iero, "Edge machine learning for AI-enabled IoT devices: A review," Sensors (Basel), vol. 20, no. 9, p. 2533, 2020.

[12]. J. Sun, C. Yang, T. Tanjo, K. Sage, and K. Aida, "Implementation of self-adaptive middleware for mobile vehicle tracking applications on edge computing," in Internet and Distributed Computing Systems, Cham: Springer International Publishing, 2018, pp. 1–15.

[13]. D. Han, N. Pan, and K.-C. Li, "A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection," IEEE Trans. Dependable Secure Comput., vol. 19, no. 1, pp. 316–327, 2022.

[14]. K. Taji, R. Ait Abdelouahid, I. Ezzahoui, and A. Marzak, "Review on architectures of aquaponic systems based on the Internet of Things and artificial intelligence: Comparative study," in The 4th International Conference on Networking, Information Systems amp Security, 2021.

[15]. H. Stewart and C. Aitken, "Prevent rather than respond: Predictive analytics for health and safety," in Day 2 Wed, September 04, 2019, 2019.

[16]. dearC, "Northstar — The Latest & Greatest in Drag-and-drop data analytics from MIT and Brown University," Towards Data Science, 04-Jul-2019. [Online]. Available: https://towardsdatascience.com/northstar-the-latest-greatest-in-drag-and-drop-data-analytics-from-mit-and-brown-university-4946dd1107cb?gi=43567d16327. [Accessed: 08-Mar-2022].

[17]. Haldorai, A. Ramu, and S. Murugan, "Signal Processing Architectures, Algorithms, and Human–Machine Interactions in Urban Applications," Computing and Communication Systems in Urban Development, pp. 49–67, 2019. doi:10.1007/978-3-030-26013-2_3

[18]. Haldorai, A. Ramu, and S. Murugan, "Artificial Intelligence and Machine Learning for Future Urban Development," Computing and Communication Systems in Urban Development, pp. 91–113, 2019. doi:10.1007/978-3-030-26013-2_5

[19]. Haldorai, A. Ramu, and S. Murugan, "Energy Efficient Network Selection for Urban Cognitive Spectrum Handovers," Computing and Communication Systems in Urban Development, pp. 115–139, 2019. doi:10.1007/978-3-030-26013-2_6

[20]. Haldorai, A. Ramu, and S. Murugan, "Social Relationship Ranking on the Smart Internet," Computing and Communication Systems in Urban Development, pp. 141–159, 2019. doi:10.1007/978-3-030-26013-2_7

[21]. Haldorai, A. Ramu, and S. Murugan, "Cognitive Radio Communication and Applications for Urban Spaces," Computing and Communication Systems in Urban Development, pp. 161–183, 2019. doi:10.1007/978-3-030-26013-2_8

[22]. Haldorai, A., Ramu, A., & Murugan, S. (2019). Machine Learning and Big Data for Smart Generation. Computing and Communication Systems in Urban Development, 185–203. doi:10.1007/978-3-030-26013-2_9.

[23]. Haldorai, A. Ramu, and S. Murugan, “Smart Sensor Networking and Green Technologies in Urban Areas,” Computing and Communication Systems in Urban Development, pp. 205–224, 2019. doi:10.1007/978-3-030-26013-2_10

[24]. G. Gokilakrishnan, S. Ganeshkumar, H. Anandakumar and M. Vigneshkumar, "A Critical Review of Production Distribution Planning Models," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 2047-2051, doi: 10.1109/ICACCS51430.2021.9441879.

[25]. S. Murugan and A. Haldorai, “Role of Machine Intelligence and Big Data in Remote Sensing,” Advances in Data Mining and Database Management, pp. 118–130, 2019.