A Survey on Low Power SBCs for Optimized Implementation of Deep Learning Models

Manjunatha Badiger
Assistant Professor, Dept. of ECE, Sahyadri College of Engineering & Management, badiger_manju@yahoo.com

Dr. Jose Alex Mathew
Professor, Dept. of AI & DS, Srinivas Institute of Technology aymanamkuzhy@gmail.com

Abstract

SBCs are low computing devices that contain all of the necessary components of a computer on a single board. SBCs can meet the goals of cost-effectiveness, minimal power usage, design and implementation flexibility and improved communication speed. Deep neural networks can benefit from the structure provided by SBCs. Deep neural network performance on SBCs can be optimized by employing efficient methods that are intended to lower the model's computational requirements. Deep neural networks on SBCs require hardware optimization as well. This may involve accelerating deep neural network processing with specialized hardware including Graphics Processing Units (GPUs) or Field Programmable Gate Arrays (FPGAs). Deep learning edge devices are becoming more and more important in a number of applications, from autonomous automobiles to industrial automation to healthcare, since they provide real-time decision-making and analysis at the edge of the network. An overview of various single board computers available on the market as well as major research efforts involved in the deployment of deep learning at the edge of computer networks is presented in this survey.

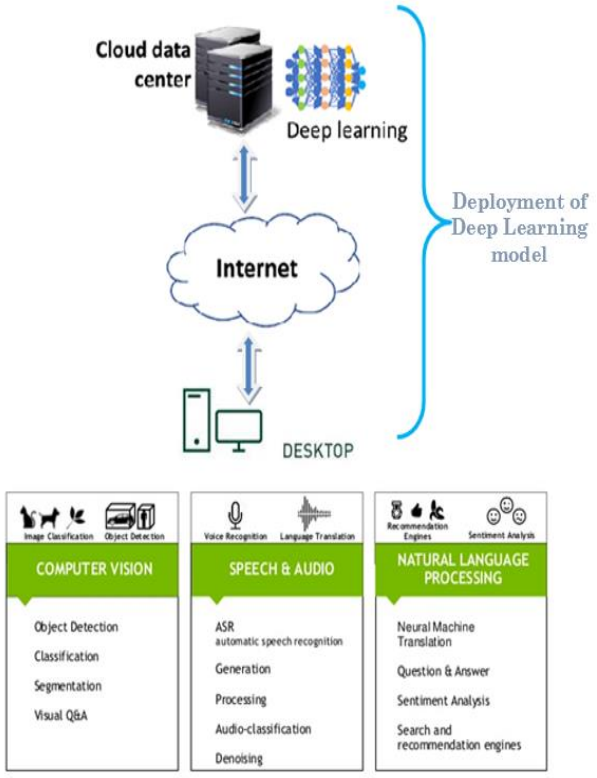
Keywords: *Single-board computers, Deep Nural Network, GPU, Edge Devices.*

I. INTRODUCTION

Deep Learning is a subset associated with machine learning and artificial intelligence (AI) which tries to mimic the manner in which individuals benefit from particular kinds of knowledge. Deep learning algorithms, as opposed to conventional device learning algorithms are piled up in growing levels of intricacy and generalization [1]. Algorithms for deep learning have successfully accomplished higher precision in a wide range of disciplines, including the processing of the image, natural language, interpretation of data, and so on. Predictive modeling and statistics are both a part of deep learning. Deep learning accelerates and makes this process simpler for data scientists, who are responsible for collecting, analyzing, and assessing enormous amounts of data [2]. In its most basic version, deep learning can be used to perform predictive analyses. Deep Learning algorithms piqued the interest of academics seeking to answer more complicated questions. For significant data mining and analysis applications, "big data" on networked IoT devices require a large number of processing resources during training and is usually shifted to central servers, which are commonly found on cloud computing platforms. There is a growing field of research in the use of convolutional neural networks (CNNs) in the recognition of images and

videos, where the network is trained to recognize patterns in the input data by sliding over the input image or video and identifying patterns in the input. In order to process sequential data, such as natural language text or speech, or time-series data, recurrent neural networks process each element in order and maintain an internal state that allows the network to remember information from previous elements in a sequential manner [3]. Generative Adversarial Networks combine two neural networks, one that generates synthetic data and the other that discriminates between real and fake data. Neural networks may be used for a variety of tasks, including data compression, anomaly detection, and picture denoising. In autoencoding, neural networks are taught to reconstitute their input data from a compressed form, known as a latent code. Research is constantly developing new architectures and techniques to improve the performance of deep learning models on a wide range of tasks in order to improve their performance, as there are many types of deep learning architectures, each with its own strengths and weaknesses. Due to network latency, obtaining the results from the cloud server can take too long for time-critical applications [4]. To fulfil the increasingly strict performance requirements of many application services, particularly in terms of latency and bandwidth, cloud computing today confronts several difficulties.

Figure 1. Cloud-Based Deep Learning Tasks



Energy efficiency, compact form factor, and cost are key requirements for Deep Learning-based smart solutions. Figure 1. depicts the system for cloud computing, which allows us to access data and applications from a distance. It is possible for users to access their data from any device in the cloud, giving them the freedom to move around without worrying about data storage restrictions [5].

Edge computing is a promising paradigm that uses single-board computers (SBCs) to improve infrastructure efficiency by providing low latency, bandwidth-efficient, cost-effective, and robust end-user services. Edge computing is the approach of performing computations as near to data sources as possible rather than in faraway, remote places. This is accomplished by an edge computing device (generally known as an edge device) deployed near the data-generating devices that have limited resources. Edge computing devices are

equipped with both processing and communication skills. Single-board computers often characterized simply as systems-on-a-chip, are extremely small computers [6]. Modern customer-driven interfaces such as High-Definition Multimedia Interface, Universal Serial Bus, Wireless Fidelity, and Bluetooth are included among them. They have SD-Cards or compact SSDs for data storage, as well as many CPU cores and RAM. SBCs have a full operating system installed, often Linux or Windows, and hence provide a wide range of general-purpose hardware. Single Board Computers are available in a variety of form factors. The most prevalent is the PC/104-compliant half-size and full-size boards. These boards may be layered on top of one another to form multi-board systems. Another prominent form factor is the 3U VPX, which is frequently utilised in military and aerospace applications.

As opposed to a computer, an SBC can offer a number of benefits which you may not be able to get from a traditional computer. One of the advantages of using them is that they consume less energy [7]. Due to the fact that they lack all of the other components required for basic operation, they are unable to perform their function properly. They also produce less heat, making them suitable for use in areas where temperature is a factor. Another advantage of Single Board Computers is their ease of use. Finally, Single Board Computers are less expensive than normal computers. The first criterion to consider while picking the best single-board computer to meet user demands. This determines what software you'll need to run, how much processing power users need, how much bandwidth users need, and a variety of other limiting considerations. SBCs are ideal for Internet of Things (IoT) applications because of their compact footprint and low power usage compared to standard desktops and laptops. The Qualities that the user should seek in an SBC are they should evaluate their priorities while purchasing the best top single-board computer for their needs. With the cheapest single-board computer devices and expensive but powerful boards, the cost is undoubtedly a consideration [8]. Size is also important. A tiny maker board for wearable technology or a robotics project, or a bigger single-board PC for sophisticated, power-hungry applications, may be what users need.

The cost comparison between edge computing and cloud-only management might change based on the unique use case and the needs of the company. Cloud-only management frequently calls for a centralised infrastructure that can deal with high data and traffic volumes. Setting up and maintaining anything like this may be expensive, especially for businesses that need a lot of data processing and storage. Contrarily, edge computing uses dispersed computing and storage resources that are positioned close to the gadgets or sensors that are producing the data. There may be a reduction in infrastructure costs as a result of this, especially for businesses that require a large number of edge devices as part of their infrastructure. Network connectivity between devices and the cloud infrastructure is crucial for cloud-only administration. This may lead to significant network expenses, especially for businesses with lots of devices producing lots of data. By processing and storing data locally instead of sending it over the network, edge computing, on the other hand, can cut down on network expenses [9]. High-performance computer resources are often needed for cloud-only management to process massive amounts of data fast.

There would be a very high cost associated with cloud computing for businesses that require complex data processing and the cost

associated with that would be very high. By placing processing resources closer to the data source, edge computing can lower processing costs by enabling quicker processing times and lowering the quantity of data that needs to be transported over the network. In order to keep the infrastructure in a safe and sound state, cloud-only management requires that it be regularly maintained and upgraded. By dispersing computer resources and simplifying the centralised infrastructure, edge computing can save maintenance costs. Ultimately, the cost comparison between edge computing and cloud-only administration will be based on the unique use case and organisational needs. In general, edge computing can result in infrastructure, network, and processing cost savings, but owing to the scattered nature of the computing resources, it may also result in greater maintenance costs.

Depending on the type of single-board computer, it can range from a small package to a larger board, with various CPUs, a variety of memory, storage options, and input/output options[10]. It is important to determine the type of SBC needed for a specific project based on its unique requirements. A number of factors have to be considered, including the amount of processing power necessary, the operating system that is needed, the accessibility of interfaces, as well as power consumption constraints. In terms of power consumption, some SBCs are designed to consume the least amount of power possible while others are designed to perform optimally in applications that require high computation speeds. In recent years, Raspberry Pi has become one of the most popular SBCs for hobbyists and educators because of its low price and simple interface. In contrast, NVIDIA Jetson is a family of SBCs designed to be used for applications requiring high levels of processing power, such as artificial intelligence and machine learning. Edge computing devices are becoming increasingly popular due to their ability to process data faster and more efficiently than traditional cloud-based solutions. By processing data at the edge, edge computing devices can reduce the latency and bandwidth requirements for transmitting data back to the cloud, thereby improving the system's overall performance. The following table 1 shows a comparison of various types of single-board computers and their specifications in the form of a comparative analysis.

TABLE I. Comparison Table of different Single Board Computer

Model	Raspberry Pi 4 Model B	BeagleBone Black	NVIDIA Jetson Nano	ASUS Tinker Board 2	ODROID-XU4
Price	$35 - $75	$55	$99 - $199	$100	$59
CPU	Broadcom BCM2711, Quad-core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz	TI Sitara AM3358BZCZA100, 1GHz ARM Cortex-A8	NVIDIA Maxwell architecture with 128 NVIDIA CUDA cores, ARM Cortex-A57 MPCore processor	Rockchip RK3399, Dual-core ARM Cortex-A72 + Quad-core ARM Cortex-A53	Samsung Exynos5422 Cortex™-A15 2Ghz & Cortex™-A7 Octa core CPUs

GPU	Broadcom VideoCore VI, OpenGL ES 3.0	PowerVR SGX530	NVIDIA Maxwell architecture with 128 NVIDIA CUDA cores, 1 TFLOPS	Mali T860MP4	ARM® Mali™-T628 MP6
RAM	2GB, 4GB or 8GB LPDDR4-3200 SDRAM	512MB DDR3	4GB 64-bit LPDDR4 1600MHz	2GB LPDDR4 3200MHz	2GB LPDDR3
Storage	MicroSD card expansion slot	4GB eMMC, MicroSD card expansion slot	MicroSD card expansion slot	16GB eMMC, MicroSD card expansion slot	eMMC module socket, MicroSD card expansion slot
Operating System	Various Linux distributions, Windows 10 IoT Core, Raspberry Pi OS	Angstrom Linux	Linux4Tegra, based on Ubuntu 18.04	Debian 10, Android 11 or latest version	Ubuntu 14.04,Andriod, Lakka
Weight	46 gram	120 grams	46 gram	55 gram	60 gram
Connectivity	Gigabit Ethernet, 2.4 GHz and 5.0 GHz IEEE 802.11ac wireless, Bluetooth 5.0, BLE, 2-USB 3.0 ports, 2-USB 2.0 ports, 2-micro-HDMI ports (up to 4kp60 supported), 1 - CSI camera port v1.4, 1-DSI display port v1.4, 40-pin GPIO header, micro-SD card	10/100 Ethernet, USB client for power and communications, USB host	Gigabit Ethernet, 2-USB 3.0 Type-A, 2-USB 2.0 Micro-B, 1-HDMI 2.0, 1-DisplayPort 1.2, MIPI CSI-2 DPHY lanes and MIPI DSI-2 DP lanes, GPIO header	Gigabit Ethernet, 2.4/5.0GHz WiFi, Bluetooth 5.0, BLE, 4-USB 2.0 ports, 1-USB 3.2 Gen1Type-A port, 1-3.5mm audio jack, 1-MIPI DSI, 1-MIPI CSI-2, 40-pin GPIO header	Gigabit Ethernet, 2-USB 3.0 host, 1-USB 2.0 host, 1-HDMI, 1-eMMC module socket, 1- microSD slot, 1-UART, 1-Infrared Rx, 1- Infrared Tx, 1-Audio Jack

II. LITERATURE REVIEW

There are a number of well-liked Single Board Computers (SBCs) on the market, each with special characteristics and functionalities. A review of a few of the most popular SBCs is provided below. Single-board computers have become highly prevalent in recent years due to their adaptability, affordability, and small size. They can be used for robotics, house automation, image analysis, and teaching. The literature analysis focuses on the benefits and drawbacks of various SBCs and offers suggestions for choosing the best SBC for a given application.

D. F. Vera, D. M. Cadena and J. M. Ramírez [11] in the research study focused on analyzing and running an iris identification algorithm on an economical and energy-efficient ARM-based board using OpenCV, a free and open-source computer vision framework. The device's viability for system application is established based on experimental results, given that some algorithm stages are upgraded to shorten the working time.

B. Dolwithayakul, P. Boonnasa, S. Klomchit and S. Tuwachaosuan [12] proposed a Raspberry Pi 2 low-power single-board computer interactive kiosk with a computer registration system, a virtual machine-based server with a RESTful web service, and integrated software on the client for turning off the inactive computer were all used . According to the study, the new registration system can reduce energy usage by up to 44.32 per cent, staff maintenance time is reduced, limited computer resources are shared equitably, and student utilization is accurately documented.

O. Haffner, E. Kučera and M. Bachúriková [13] proposed Modern techniques for assessing the integrity of welds. Visual examination is one of the joint quality assessment techniques. This research examines the viability of applying visual system-based algorithms for welding detection and assessment in single-board computers. A very useful interface for a labourer or his shift boss is an Android smartphone that can interact with this single-board computer. This article proposes a method for evaluating the quality of welds using a single-board computer, an Android smartphone, and other cutting-edge technology.

N. S. Desai and J. S. R. Alex [14] in the article proposed that carbon dioxide and carbon monoxide levels in the air are measured using pollution detection sensors along with GPS position and uploaded to Azure cloud services. Gas sensors and inexpensive integrated Beagle bone boards are used for data collection. With the help of Microsoft's Azure Machine learning tool and the use of prior data, predictive pollution metrics are acquired. The sensors gather information via calibrated gas sensors, which are subsequently transmitted to the cloud where they can be analyzed by a number of other cloud services. By avoiding pollution sources, the suggested method is put into practice and helpful for tracking and decreasing pollution in a smart city.

S. Holly, A. Wendt and M. Lechner [15] their research addresses the impacts of various CPU and GPU parameters on power usage, delay, and energy for full DNNs and individual layers, providing a measurement basis for power prediction on NVIDIA Jetson devices. Additionally, it offers the NVIDIA Jetson Nano's ideal power and energy usage options. A transition from cloud to peripheral computing is made possible by enhancing the powers of embedded devices and Deep Neural Networks (DNN) accelerators, and intelligent energy management is essential.

A. Süzen, B. Duman and B. Şen [16] compared the effectiveness associated with single-board processors in the NVIDIA Jetson Nano, NVIDIA Jetson TX2, and Raspberry PI4 boards using a CNN technique created using images of fashion items in order to compare their respective effectiveness [16]. As part of the collection, which is divided into sections such as 5K, 10K, 20K, 30K, and 45K, 45K images are contained within that collection. In order to obtain high-accuracy decisions using minimum hardware requirements in deep learning applications, it has been investigated how embedded system boards can be utilized with different data sets in the CNN algorithm.

S. Karunaratna and P. Maduranga [17] proposed that the Modern Internet of Things (IoT) advancements have made it possible for intelligent processing programs to run on small hardware components like Single Board Computers (SBC). Exciting and essential uses have been made possible by sound event detection and classification. This paper compares studies using SBC for machine learning along with deep learning-based apps to that of a standard computer and a Raspberry Pi.

B. A. M. Patel and H. R. Hamirani [18] proposed that Communications and Information Technology (CIT) can affect how well students learn. According to research, teachers are equipped with higher-order thinking skills because they are taught to use digital devices in the classroom and are technologically proficient. India requires cost-effective and cutting-edge educational equipment to support STEM instruction. We created affordable educational laptops for teachers, pupils, and institutions with strong connectivity and computing capacity. The physical computing feature will improve the young students' computer abilities and inventiveness.

Z. Ozkan, E. Bayhan, M. Namdar and A. Basgumus [19] their research evaluate the techniques for deep learning-based danger identification and recognition, assessed in terms of the military and defence sector, using autonomous aerial vehicles and the Raspberry Pi platform. (UAV). Convolutional neural networks, one of the deep learning methods, are used for object-based machine learning training. The model, which is then moved to the Raspberry Pi 4 Model B electronic device, is trained using data sets with pictures chosen from various weather, land conditions, and times of the day. With the help of pictures and data from military operations captured by UAVs using the Raspberry Pi Camera V2 module, the technique for identifying and detecting things is evaluated. In terms of object identification and recognition accuracy, the Faster-RCNN design has obtained a rate of %91 while the SSD MobileNet V2 architecture has only seen %88.

M. I. Uddin, M. S. Alamgir, M. M. Rahman, M. S. Bhuiyan and M. A. Mora [20] in the article proposed a novel method for dealing with transportation congestion. It will monitor real-time vehicle traffic and make decisions based on the circumstance. It can also identify the license plate of cars that break traffic laws and transmit a proof notification to the driver. The document tries to depict the advancement and efficiency of this system over the current manual traffic management systems.

U. Farooq, A. Rehman, T. Khanam, A. Amtullah, M. A. Bou-Rabee and M. Tariq [21] in the research makes the utilization of YOLOv4- tiny to address the issue to handle the problem of maintaining a balance between cost-effectiveness and algorithm efficiency, this study employs YOLOv4- tiny. Farmers now have access to an affordable IoT weed detection option thanks to this powerful and efficient fast deep learning approach, which is able to operate on devices that have reduced processing power. The research describes how to use a lightweight model that they trained on a Raspberry Pi 4 Model B and suggests using YOLOv4-tiny, an object recognition model, for weed identification.

Kaspars Sudars,Ivars Namatevs,Janis Judvaitis,Rihards Balass and Arturs Ņikuļins [22] in this research presented a deep neural network (DNN) approach for raspberry as well as quince recognition on RGB pictures. It is gleaned from the YOLOv5 architecture and has seven berry classes corresponding to different

stages of berry growth. It is feasible to obtain an overall accuracy of approximately 80.9% and, in certain instances, as high as close to 95% for some classifications using DNN architecture. The DNN model, which was trained on labelled data gathered during the AKFEN project, is publicly available on the GIT repository. High-throughput phenotype determination necessitates more investigation into Sensor Networks, Wireless Systems, 3D point cloud processing, and multi-spectral picture processing.

A. Jamini, S. S. L. Perubhotla, G. R. Patlola, S. Velpula and S. G.L discussed that in 2019 Coronavirus Disease (COVID-19) has claimed the lives of many people and wreaked havoc on society. Many people have felt anxious, particularly those who are accustomed to being outside and engaging in lots of social contacts. A facial shield and hand sanitizer can help prevent the spread of the infection and offer protection. The well-being of a person can also be strongly inferred from their bodily temperature. This study presents a deep learning system using open cv that can be used to identify situations in which hoods aren't worn and people whose body temperatures are higher than usual. A smartphone program that gets text messages and a sanitizer dispenser can be used to take additional measures[23]. This model efficiently keeps an eye on people, particularly in crowded places like theatres, colleges, and workplaces.

J. Bogacz and A. Qouneh [24] in this research proposed that Beaglebone Black Wireless is used to build and characterize a Convolutional Neural Network (CNN). The two case studies being tested are facial recognition and hand-drawn digits. To reduce reaction time and data usage, machine learning (ML) processing can be shifted to the Beaglebone, a cloud network peripheral device. Describe the execution times and battery consumption for the regular BeagleBone Black and the AI version.

Dharshini S,Haneesh T,Venugopal E, Rama Devi S, Sree Dhviya M and P. Sivakumar [25] in the article investigate the Single Board Computer (SBC) configuration and to construct a rudimentary application for exchanging emails utilizing Simple Mail Transfer Protocol (SMTP). The BeagleBone Black Wireless board, which was used to install the OS on an SD card, serves as the research's basis. In addition, the Wi-Fi has been configured up for connection to the network, and the GPIO has been organized on the board. The external LED is activated to act as an indication to show that the receiver has received communications.

M. Slabinoha, S. Melnychuk, I. Manuliak and B. Pashkovskyi [26] in the article proposed SQLite, LMDB, and UnQLite embedded databases for single-board computers are compared to NoSQL embedded databases in terms of their performance. This document compares database link Python software's performance for CRUD operations. The hardware and software for the exercise were selected, and the findings were acquired[27]. The findings can be applied to the development of Python software for single-board processors with significant amounts of data in the device's local memory.

P. S. Christopherson, A. Eleyan, T. Bejaoui and M. Jazzar [27] their research article seeks to investigate an affordable, portable assistive technology that improves the quality of life for its user. In order to assist the visually disabled in making decisions, it makes use of smart technology like sensors, cameras, and vocal commands to help them spot and avoid possible hazards as they travel . The primary microcontroller is a Raspberry Pi, which serves as the interface for all the sensors and houses a

deep learning model that enables the classification of people within 1.2 meters of the visually disabled individual.

T. S. Mohammed and O. A. L. A. Ridha [28] in the article they proposesed a DenseNet-201 pre-trained network-based method for the COVID-19 detection in Chest X-Ray pictures [29]. A Raspberry Pi4 equipped with an 8MP camera and connected to a PC powers the device. Testing the system on 20 distinct pictures yielded an accuracy of 99.2% and an AUC of 99.92%. All x-ray images were properly categorized.

O. Simonoski, I. Mitreska and I. Dimitrievski [29] in article they proposed to develop an extremely precise and real-time method for detecting non-mask wearers in order to compel them to put on masks to be able to help the community's health. It employs a collection of images with and without filters to recognize features in a live video relayed via the Camera module using OpenCV, and it then transmits the data to the cloud for observation and additional analysis. The complete process is carried out on a Raspberry Pi in order to ensure it simple, quick, expandable, and efficient. It uses OpenCV to identify features in real-time from a live stream via the Camera module using a set of pictures with and without masks and then sends the data to the cloud for viewing and further analysis. The entire procedure is run on a Raspberry Pi to make it accessible, fast, scalable, and effective. This initiative allows for greater control over the information that has already been given and emphasizes the use of our technique to prevent local infection and reduce the chance of human coronavirus disease carriers.

III. DEEP LEARNING ON THE CUTTING EDGE

There are a number of constraints that have to be considered when optimizing for deep learning. When deep learning is applied at the periphery, it can address the aforementioned issues in addition to receiving other benefits. The edge refers to the specific computation performed on consumer products. The term "edge" in this context refers to the specific computation carried out on consumer products. Many compelling reasons should be given as to why edge computing should be chosen over cloud computing in the future.

1. Bandwidth and Latency

Bandwidth can be measured by looking at how much data is delivered over a network for a given period of time in addition to how much data is delivered over a given distance. The term network latency, also referred to as lag, is a term used to characterize delays in network communication that are caused by network latency. Essentially, it is more beneficial to think about it in terms of networking as the amount of time it takes for a packet of data to pass from its source to its destination and then be decoded once it has reached its destination. A network is referred to be low-latency when transmission delays are minimal and as high-latency when transmission delays are greater. High latency causes bottlenecks and is linked to poor quality of service, jitter, as well as a bad consumer experience. The impact of latency can be either transient or permanent, depending on the cause of the delays. Transmission delays can be excessively long in cloud computing. Many new types of applications have stringent delay requirements that the cloud would struggle to deliver on a constant basis. Transmission delays can be excessively prolonged in cloud computing. Many new

types of applications have stringent delay requirements that the cloud would struggle to deliver on a constant basis.

2. Security and Decentralization

Commercial servers are vulnerable to attacks and breaches, making security a major issue in cloud computing. Of course, if the vendor is reliable, the risk is minimal. You must rely on a third party to protect your intellectual property and data. When an edge computing device is connected to the internet, the IP address can be completely controlled. Decentralization is the availability of cloud computing resources across decentralized computer nodes that create a peer-to-peer network in which there is no one body with centralized control. However, having a number of edge devices allows for full decentralization benefits.

3. Customization

A customized SBC is a ready-to-use embedded platform comprised of a carrier board and a computer on a module. The platform may be expanded to meet future demands by upgrading to another pin-compatible Computer on a Module based on the most recent CPUs. There is no question that this platform may be viewed as an adaptable platform since CPUs, memory, and I/O within the platform's framework can be picked based on the application's requirements, which is critical to remember while picking CPUs, memory, and I/O in connection to the platform.

4. Swarm Intelligence

There are a number of ways in which machine learning models can be trained via edge devices. In particular, this is particularly useful when it comes to Reinforcement Learning since it allows you to simulate a large number of "episodes" at the same time. Data can also be collected from edge devices in order to be used for Online Learning (or Continuous Learning). A single model may be trained in parallel across all edge devices using optimization approaches such as Asynchronous SGD. It may also simply be used to aggregate and process data from numerous sources.

5. Redundancy

A durable memory and network design must have redundancy. A network's other nodes may suffer significant consequences if one node fails. Edge devices can offer an adequate amount of redundancy. A node's neighbour can temporarily take over if one of the edge devices malfunctions. This substantially increases dependability and significantly cuts downtime.

6. Financial sustainability over the long term

Cloud-based services will eventually be more expensive than having a specialised set of inference devices. Furthermore, edge devices, which are less expensive to create in mass, will eventually prove to be more expensive than regular versions. Additionally, the duty cycle for inferencing devices is large when they are continually operating, which implies that cloud-based solutions often cost more over time.

IV. DEEP LEARNING'S CONSTRAINTS ON THE EDGE

Deep Learning models are frequently larger and more computationally expensive than other machine learning algorithms. This is partly due to their reliance on large amounts of data, which can be challenging to store and process within edge devices. Several approaches have been proposed in order not only to accommodate these deep learning models but also to reduce their size and computational cost.

1. Parameter Efficient Neural Networks

The massive scale of neural networks is a gaze aspect. Because edge devices are often incapable of handling vast networks, researchers were encouraged to design parameter efficient models while preserving accuracy. SqueezeNet and MobileNet are two prominent efficient neural network topologies. SqueezeNet combines late-down sampling and filter count reduction methods with "Firemodules" that execute shallow and deep 1x1 convolutions. This leads to good performance with a small number of parameters. A similar arrangement is used by the MobileNet, but with a deep 1x3 convolution.

2. Pruning and Truncation

A substantial percentage of neurons in trained networks are benign and contribute nothing to ultimate accuracy. We can prune such neurons to conserve space in this situation. Google's Learn2Compress discovered that we can reduce the size by a factor of two while preserving 97% of the accuracy. Furthermore, the majority of neural network parameters are 32-bit float values. Edge devices, on the other hand, maybe engineered to function with values as low as 8 bits. Reducing accuracy can result in a large reduction in model size. For example, downsizing a 32-bit model to an 8-bit model decreases the model size by a factor of four.

3. Distillation

Distillation is an effective method of teaching smaller networks by taking advantage of the experience from bigger "teacher" networks. Google's Learn2Compress approach uses this principle for size reduction and provides a powerful tool for reducing the model size without compromising accuracy. This enhancement makes it a great option for businesses looking to reduce the size of a model while still maintaining model accuracy.

4. Optimized Microprocessor Designs

There have been many discussions about how to reduce neural networks in order to incorporate peripheral devices into a smaller form factor. It would also be possible to scale up the performance of the microprocessor as an alternative method (or as a complementary method). An easy solution to this problem would be to use a GPU on a microprocessor, such as the popular Nvidia Jetson, which has a GPU built into it. Despite this, if these devices are deployed on a large scale, they may not be cost-effective to deploy. As an alternative solution, Vision Processing Units (VPUs) might be of interest. The Movidius VPU is claimed by Intel to be the fastest VPU on the market today due to its ultra-low power consumption and high performance. In addition to Google's AIY kits, Intel's Neural Compute Stick also utilizes this kind of VPU internally. As an alternative, we could use FPGAs since they consume less power than GPUs and can accommodate lower-bit (< 32-bit) architectures. In spite of this, there is likely to be a slight drop in performance compared to GPUs due to their lower FLOPs rating compared to GPUs. The best way to deploy custom ASICs on a large scale would be to use them for large-scale deployments. A microarchitecture similar to NVIDIA's V100 could be fabricated to accelerate matrix multiplication in a way that could dramatically improve performance. The term "optimised microprocessor designs" describes the application of cutting-edge methods and tools to improve the performance, power usage, and general effectiveness of microprocessors. The majority of contemporary computing devices, from smartphones to supercomputers, are built

around microprocessors, and the effectiveness of these components depends on how well they work. Engineers employ a number of methods, including:

Architecture optimization: This involves optimising speed while reducing power usage in the microprocessor layout. Techniques like pipelining, branch prediction, and out-of-order processing may be included in this.

Transistor scaling: Reducing the size of the microprocessor's transistors can boost speed and lower power usage. There are restrictions on how small transistors can be produced, but novel technologies like gate-all-around transistors and FinFETs are being created to overcome these restrictions.

Low power design: There are several techniques that microprocessors can use to reduce their power consumption, such as clock gating, power gating, dynamic voltage scaling, and frequency scaling.

Memory optimization: The architecture of the memory hierarchy can be optimised to increase speed. Memory access is a major performance bottleneck for microprocessors.

Parallelism: Methods like multi-core processing, hyper-threading, and vector processing can be used to create microprocessors that can handle numerous jobs concurrently.

Customization: By designing customized microprocessors for particular uses, speed, and power consumption can be increased.

There are many ways to scale down neural networks to fit into edge devices, but there is another complementary method that can be used to improve the performance of a microprocessor. The simplest solution would be to include a GPU in the microprocessor, such as the popular Nvidia Jetson. These GPUs are typically significantly more powerful than the built-in processors on most devices and can be used to drastically improve the speed of a neural network. The added benefit of this is that it requires no additional coding and can be done simply by adding a GPU. Including a GPU on a microprocessor allows for more complex tasks such as image detection and object recognition to be achieved by edge devices. This increases the performance of the device, allowing for more complex tasks to be done in less time. Furthermore, since GPUs are powerful enough to handle these types of tasks, they can also help with the training of neural networks which may result in improved efficiency of the system. Additionally, having a GPU on a microprocessor enables edge devices to run multiple neural networks simultaneously and hence offer better performance compared to single neural network implementations on traditional processors. Therefore, having a GPU on a microprocessor is an effective way of scaling up the performance of edge devices and helping them achieve more.

V. ADVANTAGES AND DISADVANTAGES OF SBCS

SBC advantages

An SBC is a complete computer that possesses every requirement for a functional computer. The functioning of different devices is frequently made possible by a tiny working computer on a single integrated circuit. Due to its versatility and advantages over desktop computers, it may be used in a variety of industries.

Easy to Use: SBCs are simple and compact, in contrast to desktop computers, which can be difficult for certain users to assemble. These single computer units are simple to use because they all include a variety of built-in

functionality. As a result, users find it straightforward and easy to implement in a variety of industries.

Verified Hardware: The hardware has undergone numerous modifications and enhancements since SBCs first appeared, allowing it to be used. Developers have also minimized lag time and other issues that may cause customers to be cautious to explore these PCs. Utilizing SBCs appears to be the easiest solution because the associated risks are kept under control.

Adaptable: An SBC allows users to make changes to their computers, rendering it a versatile tool that can be used for many different tasks. In this way, it becomes a dynamic feature that is able to meet the needs of every user without having to purchase a new computer system to meet those needs. It also works effectively with a variety of devices from multiple sectors, which makes it a useful tool for a variety of users.

Single Source: Single Source Single board systems, unlike system-on-a-chip technologies, operate independently as long as they are designed to perform a specific function. It allows users to do several functions on a single-board computer.

Time to Market: SBCs are not only quick and effective; their design and functionality also make them suited for a variety of uses. Being adaptable and having no downtime are further advantages. They are well-liked by consumers because of these and other advantages. Additionally, the design is quicker than other computer types.

SBC disadvantages

These computers have drawbacks while being superior to standard computing equipment in terms of benefits. However, its drawbacks focus on usability rather than functionality. Here are a few single-board computer systems' drawbacks.

Cost: These expensive computers also cost more to manufacture due to their complex design. As a result, SBCs are sometimes more expensive than regular computers and difficult for some customers to get. This is a problem, particularly for customers on a tight budget who want to utilise SBCs for a variety of tasks.

Flexibility: A system-on-a-chip is usually more beneficial when it comes to customisation than single-board computing systems. While SBCs have an adaptive characteristic, other computer systems frequently have greater flexibility than SBCs. This makes the computer unpredictable, which is exacerbated when trying to modify it for specific tasks.

Knowledge: Users must spend some time learning about single-board computer systems' creation and functionality before utilising them. This is due to the fact that it differs greatly from other computing devices and can be challenging to use without a clear grasp of how it works. Some consumers may find it difficult to utilise the item after purchasing it if they use such a tactic.

VI. APPLICATIONS OF SBCS

Similar to other computing devices, computers constructed from a single board frequently carry out the same functions. They do, however, have several benefits over frequently used devices, making them appropriate for a variety of applications. Due to their effectiveness, lack of downtime, and ability to do several tasks simultaneously, SBCs are now considered vital in the majority of businesses. Following that, the following are some of the uses for SBCs:

Experimentation: SBCs have established themselves as the ideal computer platforms for carrying out numerous scientific tasks. Such a method has sparked the development of fresh findings about many topics.

Learning Programming: SBCs make it simple for computer students to master computer coding procedures regardless of the programming language they use. This makes it possible for students to swiftly acquire and understand the most important study ideas.

Creation of Media Players: Using this sort of computer, users may design a good media player due to its capacity, input/output functions, and adaptability.

Automotive: The majority of contemporary automobiles employ single-board computers to manage sensors, give precise map data, and other functions. These days, several car manufacturers employ this as a standard feature.

Robotics: As they enable the fusion of science, engineering, and technology, SBCs play a crucial role in delivering necessary data for robotics. This eventually leads to the development of moving machines that carry out particular functions.

Performing Computing Tasks: Single-board computers may also carry out common computing activities that are handled by desktops and laptops. These consist of text processing and online browsing.

Facilitates IoT: SBCs are essential to the Internet of Things (IoT) industry since they make a variety of tasks easier.

VII. CONCLUSION

Deep learning models can now be optimised to run on low-power single-board computers (SBCs), demonstrating that there is a growing interest in developing energy-efficient deep learning hardware solutions as a consequence of optimised deep learning model execution on SBCs. A critical driving factor for the implementation of deep learning models on edge devices such as smartphones, drones, and Internet of Things devices is the need to optimize their power usage. The survey suggests using low-power SBCs like the Raspberry Pi, Jetson Nano, Coral Dev Board, and UP Squared AI Vision X Developer Kit to run deep learning models. Each of these SBCs has unique benefits and drawbacks in terms of performance, energy use, and software support. There are many optimisation methods, such as model reduction, quantization, and pruning, to lower the power requirements of deep learning models. With these methods, deep learning models' processing needs can be drastically reduced without losing any of their precision. Overall, the survey has demonstrated that low-power SBCs can be an effective choice for deploying deep learning models on peripheral devices with constrained power supplies. To enhance the functionality and energy economy of these devices for deep learning apps, more research and development are still needed.

REFERENCES

[1] S. V. Mahadevkar et al., "A Review on Machine Learning Styles in Computer Vision—Techniques and Future Directions," in IEEE Access, vol. 10, pp. 107293-107329, 2022, doi: 10.1109/ACCESS.2022.3209825.

[2] Mahmoud Abbasi et al., “Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey”, Computer Communications, Volume 170, 15 March 2021, Pages 19-41, doi:10.1016/j.comcom.2021.01.021

[3] N. M. Rezk, M. Purnaprajna, T. Nordström

and Z. Ul-Abdin, "Recurrent Neural Networks: An Embedded Computing Perspective," in IEEE Access, vol. 8, pp. 57967-57996, 2020, doi: 10.1109/ACCESS.2020.2982416.

[4] W. Yu et al., "A Survey on the Edge Computing for the Internet of Things," in IEEE Access, vol. 6, pp. 6900-6919, 2018, doi: 10.1109/ACCESS.2017.2778504.

[5] W. Yu et al., "A Survey on the Edge Computing for the Internet of Things," in IEEE Access, vol. 6, pp. 6900-6919, 2018, doi: 10.1109/ACCESS.2017.2778504.

[6] C. Li, J. Chen and M. Jiang, "Research on Application of Embedded Single Chip Computer Intelligent Control," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020, pp. 588-591, doi: 10.1109/ICMCCE51767.2020.00132.

[7] M. H. Chowdhury, S. Shuvro Mistry, S. N. Shamaem, O. Numan and R. A. Tuhin, "Lowering the Power Consumption of Healthcare Centers in Bangladesh using Low-Power Consuming Computing Devices," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 1110-1115, doi: 10.1109/ICCIT57492.2022.10054750.

[8] S. J. Johnston, M. Apetroaie-Cristea, M. Scott and S. J. Cox, "Applicability of commodity, low cost, single board computers for Internet of Things devices," 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, USA, 2016, pp. 141-146, doi: 10.1109/WF-IoT.2016.7845414.

[9] K. Cao, Y. Liu, G. Meng and Q. Sun, "An Overview on Edge Computing Research," in IEEE Access, vol. 8, pp. 85714-85728, 2020, doi: 10.1109/ACCESS.2020.2991734.

[10] E. Lee, H. Oh and D. Park, "Big Data Processing on Single Board Computer Clusters: Exploring Challenges and Possibilities," in IEEE Access, vol. 9, pp. 142551-142565, 2021, doi: 10.1109/ACCESS.2021.3120660.

[11] D. F. Vera, D. M. Cadena and J. M. Ramírez, "Iris recognition algorithm on BeagleBone Black," 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Warsaw, Poland, 2015, pp. 282-286, doi: 10.1109/IDAACS.2015.7340744.

[12] B. Dolwithayakul, P. Boonnasa, S. Klomchit and S. Tuwachaosuan, "Green public computer lab using single-board computer and interactive computer reservation system," 2015 International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2015, pp. 1-4, doi: 10.1109/ICSEC.2015.7401420.

[13] O. Haffner, E. Kučera and M. Bachúriková, "Proposal of weld inspection system with single-board computer and Android smartphone," 2016 Cybernetics & Informatics (K&I), Levoca, Slovakia, 2016, pp. 1-5, doi: 10.1109/CYBERI.2016.7438600.

[14] N. S. Desai and J. S. R. Alex, "IoT based air pollution monitoring and predictor system on Beagle bone black," 2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2), Chennai, India, 2017, pp. 367-370, doi: 10.1109/ICNETS2.2017.8067962.

[15] S. Holly, A. Wendt and M. Lechner, "Profiling Energy Consumption of Deep Neural Networks on NVIDIA Jetson Nano," 2020 11th International Green and Sustainable Computing Workshops (IGSC), Pullman, WA, USA, 2020, pp. 1-6, doi: 10.1109/IGSC51522.2020.9290876.

[16] A. A. Süzen, B. Duman and B. Şen, "Benchmark Analysis of Jetson TX2, Jetson Nano and Raspberry PI using Deep-CNN," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2020, pp. 1-5, doi: 10.1109/HORA49412.2020.9152915.

[17] S. Karunaratna and P. Maduranga, "Artificial Intelligence on Single Board Computers: An Experiment on Sound Event Classification," 2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), Colombo, Sri Lanka, 2021, pp. 1-5, doi: 10.1109/SLAAI-ICAI54477.2021.9664746.

[18] B. A. M. Patel and H. R. Hamirani, "Affordable Educational Laptop With Physical Computing Using Single Board Computer," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-4, doi: 10.1109/ICCCI50826.2021.9402691.

[19] Z. Ozkan, E. Bayhan, M. Namdar and A. Basgumus, "Object Detection and Recognition of Unmanned Aerial Vehicles Using Raspberry Pi Platform," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2021, pp. 467-472, doi: 10.1109/ISMSIT52890.2021.9604698.

[20] M. I. Uddin, M. S. Alamgir, M. M. Rahman, M. S. Bhuiyan and M. A. Moral, "AI Traffic Control System Based on Deepstream and IoT Using NVIDIA Jetson Nano," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh, 2021, pp. 115-119, doi: 10.1109/ICREST51555.2021.9331256.

[21] U. Farooq, A. Rehman, T. Khanam, A. Amtullah, M. A. Bou-Rabee and M. Tariq, "Lightweight Deep Learning Model for Weed Detection for IoT Devices," 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), Patna, India, 2022, pp. 1-5, doi: 10.1109/ICEFEET51821.2022.9847812.

[22] K. Sudars et al., "YOLOv5 Deep Neural Network for Quince and Raspberry Detection on RGB Images," 2022 Workshop on Microwave Theory and Techniques in Wireless Communications (MTTW), Riga, Latvia, 2022, pp. 19-22, doi: 10.1109/MTTW56973.2022.9942550.

[23] A. Jamini, S. S. L. Perubhotla, G. R. Patlola, S. Velpula and S. G.L, "Face Mask and Temperature Detection for Covid Safety using IoT and Deep Learning," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 373-379, doi: 10.1109/ICCES54183.2022.9835768.

[24] J. Bogacz and A. Qouneh, "Convolution Neural Network on BeagleBone Black Wireless for Machine Learning Applications," 2022 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2022, pp. 1-4, doi:

10.1109/URTC56832.2022.10002216.

[25] Dharshini S,Haneesh T,Venugopal E, Rama Devi S, Sree Dhviya M and P. Sivakumar, "Implementing BeagleBone Black as a Single Board Computer by Transferring E-mail using SMTP," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 1184-1187, doi: 10.1109/ICACRS55517.2022.10029224.

[26] M. Slabinoha, S. Melnychuk, I. Manuliak and B. Pashkovskyi, "Comparative analysis of embedded databases performance on single board computer systems using Python," 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2022, pp. 222-225, doi: 10.1109/CSIT56902.2022.10000475.

[27] P. S. Christopherson, A. Eleyan, T. Bejaoui and M. Jazzar, "Smart Stick for Visually Impaired People using Raspberry Pi with Deep Learning," 2022 International Conference on Smart Applications, Communications and Networking (SmartNets), Palapye, Botswana, 2022, pp. 1-6, doi: 10.1109/SmartNets55823.2022.9993994.

[28] T. S. Mohammed and O. A. L. A. Ridha, "Implementation of Deep Learning In Detection of Covid-19 In X-ray Images Using Raspberry Pi," 2022 Iraqi International Conference on Communication and Information Technologies (IICCIT), Basrah, Iraq, 2022, pp. 203-208, doi: 10.1109/IICCIT55816.2022.10010353.

[29] O. Simonoski, I. Mitreska and I. Dimitrievski, "Intelligent Real-Time Face Mask And Temperature Recognition Using Deep Learning On Raspberry Pi," 2022 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2022, pp. 164-168, doi: 10.1109/ICAI55857.2022.9960126.