Analysis the Principle of AI Chip Principle and the State-of-art Applications

Derun Li

Department of Electronic Engineering, University of Tennessee, Knoville, United States

Dli39@vols.utk.edu

Abstract. In recent years, AI chips are widely used in advance IC manufacturing. On this basis, this study provides an in-depth analysis of the principles and state-of-the-art applications of AI chips. This research first briefly introduces the history of artificial intelligence and the emergence of artificial intelligence chips as specialized hardware components aimed at effectively performing complex calculations required for artificial intelligence tasks. The article then delves into the basic definition and description of AI chips, the differences between AI chips and ordinary chips, the architecture and manufacturing process of AI chips, application scenarios and results, and the limitations of current AI chips. At the same time, the study also emphasizes the impact of artificial intelligence chips on the field of artificial intelligence research and their potential to shape the future of computing. Overall, the result of the study is a valuable resource for anyone interested in understanding the principles and applications of AI chips.

Keywords: AI chips, artificial intelligence, architecture, manufacturing process.

1. Introduction

The development of artificial intelligence has a long and fascinating history, which can even be traced back to ancient myths and philosophical discussions about artificial organisms. The concept of modern artificial intelligence emerged during World War II, when Alan Turing proposed the idea of a universal machine that could perform any calculation. Turing discussed in the article how to build intelligent machines and how to test their intelligence [1]. The development of AI chips (also known as AI accelerators or AI processing units (APUs)) has played an important role in advancing the field of artificial intelligence. These specialized hardware components are designed to efficiently perform complex calculations required for artificial intelligence tasks. In 1986, Hinton co authored a paper titled Learning presentations by back propagation errors [2]. The algorithm called back propagation that appears inside can significantly improve the performance of multi-layer or "deep" neural networks. With the progress of artificial intelligence research, there is an increasing need for more efficient hardware to handle the computational requirements of artificial intelligence algorithms. Researchers began experimenting with specialized hardware, such as digital signal processors (DSPs), graphics processing units (GPUs), and field programmable gate arrays (FPGAs), to accelerate certain artificial intelligence tasks. The development of AI chips continued in the 2020's, with a focus on improving computing power, energy efficiency, and storage capacity. Introduced features such as mixed precision computing, sparsity, and novel memory architecture to further improve the performance of artificial intelligence chips. The 2020 era witnessed efforts to expand AI chips for various applications and deployment scenarios. In addition, one also explored Heterogeneous computing methods, such as combining CPU, GPU and dedicated AI chips, to achieve the best performance and efficiency of complex AI workloads.

The development of AI chip technology has been a dynamic and rapidly developing field in recent years. The progress of AI algorithms and the growing demand for more efficient and powerful AI processing have driven significant progress in AI chip development Since 2010, the breakthrough in deep learning and the success of neural networks in various artificial intelligence applications have sparked a wave of interest in developing specialized artificial intelligence chips. The rise of deep learning is closely linked to the proliferation of artificial intelligence chip startups, to the extent that many startups and established technology companies have begun to focus on designing and

manufacturing AI specialized chips to meet the growing demand for AI acceleration. Deep learning is a subfield of machine learning that uses multi-layer artificial neural networks (deep neural networks) to model and solve complex problems [3]. The rise of deep learning has sparked a demand for more efficient hardware to handle the computational requirements of training and deploying deep neural networks. This demand, coupled with the potential market opportunities brought by the constantly developing artificial intelligence industry, has led to a surge in artificial intelligence chip startups since the 2010s. AI chip technology, as a specialized hardware for optimizing AI workloads, has undergone rapid development in recent years. In the past few years, the development of special AI accelerator (also known as artificial intelligence chips or artificial intelligence processing units (APUs)) has surged. These chips are designed to perform AI computing efficiently and are tailored to specific AI workloads such as deep learning, Natural language processing, and computer vision. Special AI accelerators have been optimized for specific AI workloads, making them more efficient and energy-efficient than traditional general-purpose CPUs and GPUs. They aim to utilize parallel processing and specialized hardware architectures to accelerate matrix operations and calculations commonly found in neural networks. The AI accelerator is equipped with a customized architecture designed to maximize the performance of AI tasks. These architectures typically include dedicated hardware units for Matrix multiplication, vector operations, and other computations common in deep learning. Gupta introduces accelerators for artificial intelligence systems, which provide various human-machine interactions such as facial recognition systems, early detection of cancer [4]. To achieve these highly accurate applications, complex calculations are required, and the hardware requirements are extremely challenging. Therefore, the AI accelerator chip can enable various AI devices to complete various complex machine learning algorithms in a very short time. In general, the special AI accelerator is a key component of AI progress and deployment. Their professional architecture and efficiency improvements play an important role in driving innovation driven by artificial intelligence and achieving practical applications across industries.

With the rise of edge computing and the Internet of Things (IoT), the development of AI chips that can directly perform AI computing on Edge device has become the focus of attention. Edge artificial intelligence chips can perform real-time processing and decision-making on the device itself, reducing the need to send data to central servers. This method can enhance privacy, reduce latency, and optimize network bandwidth. As mentioned in Ref. [5], edge chips are widely used in current AI applications, not only in high-performance automotive and robot applications, but also in smartphones, IoT, wearable devices, and forever online products. At present, edge AI chips have been widely used in various industries, including smart home devices, monitoring systems, autonomous vehicle, Industrial internet of things, medical care, agriculture, etc. With the growing demand for edge computing and AI, the development of more complex and efficient edge AI chips is expected to expand the ability and possibility of edge applications driven by AI. This article mainly introduces artificial intelligence chips and their principles and application status. Provide a detailed description of the basic definition and description of AI chips, the differences between AI chips and regular chips, the architecture and manufacturing process of AI chips, application scenarios and results, and the limitations of current AI chips.

2. Basic Descriptions

AI chips, also known as artificial intelligence chips or AI processors, are specialized hardware designed to efficiently and quickly perform tasks related to artificial intelligence (AI) and machine learning (ML). These chips have been specially optimized to meet the computational needs of artificial intelligence algorithms, enabling faster and more energy-efficient processing compared to traditional general-purpose processors. Artificial intelligence chips aim to accelerate various artificial intelligence tasks, such as data processing, pattern recognition, neural network training, and inference. They are particularly suitable for tasks involving large-scale matrix operations and parallel computing commonly found in deep learning algorithms. The architecture of AI chips has been specially

designed to leverage the inherent parallelism and data flow patterns of AI workloads, significantly improving performance and energy efficiency. In the journal "The AI Chip Race" [6], it is mentioned that AI chips not only include graphics processing units (GPUs) originally designed for graphics rendering, but also specialized integrated circuits (ASICs) that have become a major component of artificial intelligence computing due to their parallel processing capabilities. They are tailored for specific artificial intelligence workloads or applications and provide excellent performance and energy efficiency by optimizing their design for a specific set of tasks, It also includes Field Programmable Gate Arrays (FPGAs).

Although not mentioned in this journal, it can be seen that AI chip also has tensor processing units (TPUs) developed by Google [7]. TPU aims to accelerate neural network computing, especially inference tasks. They excel in handling tensor operations, which are the foundation of many artificial intelligence algorithms. In addition to classifying AI chips based on their construction, it can be seen from previsou study that AI chips can also be classified into Training Chips and Inference Chips based on their purpose of use [8]. Training chips typically have higher computing power and memory bandwidth, which can effectively handle large datasets and complex model architectures. The inference chip is designed to deploy trained artificial intelligence models in real-world applications.

The training chip aims to significantly accelerate the training process of deep neural networks and other machine learning models. Training a neural network involves iteratively adjusting the weights and biases of network parameters to minimize the difference between the predicted output and the actual target value. So the main characteristics and characteristics of training chips include high computing power, high memory bandwidth, large memory capacity, and the ability to support various levels of numerical accuracy and scalability. Training chips are crucial for accelerating the development and deployment of state-of-the-art artificial intelligence models, as they provide the computing power and efficiency required to handle the enormous computational demands of deep neural network training.

The inference chip is a specialized hardware designed to deploy and execute trained machine learning models, enabling real-time prediction and decision-making in various applications. Unlike training chips that focus on optimizing the training process of neural networks, inference chips aim to efficiently process input data through pre trained models to generate fast and accurate predictions. For the design of inference chips, it is necessary to mainly consider low latency, energy efficiency, and deployment flexibility. The voice assistant, image recognition, and other artificial intelligence functions currently used on smartphones provide support, and autonomous vehicles are all applications of inference chips. With the continuous integration of artificial intelligence into various applications, the development of efficient and professional inference chips is crucial for achieving real-time decision-making and intelligent automation in various industries.

3. Principle

The principle of AI chips involves specialized hardware and architecture design, aimed at accelerating the execution of artificial intelligence (AI) tasks, especially those involving neural networks and machine learning algorithms. These principles aim to optimize performance, energy efficiency, and parallel processing capabilities. Therefore, the key principles of AI chips include parallelism, customized hardware accelerators, data flow optimization, memory computing, efficient activation functions, real-time inference, and scalability. These principles collectively lead to AI chips being able to perform AI tasks faster and more efficiently, with lower power consumption.

Parallelism is a fundamental concepts in artificial intelligence chip design, playing a crucial role in accelerating the execution of artificial intelligence (AI) tasks. These concepts have been optimized to meet the enormous computational needs of artificial intelligence algorithms, especially those involving neural networks and deep learning models. Due to the fact that artificial intelligence tasks typically involve a large amount of data processing. The AI chip responsible for executing tasks utilizes a Single Instruction Multiple Data (SIMD) architecture, while performing the same operations

on multiple data points, significantly accelerating computing speed. In order to accelerate the overall computing speed, parallelism will be considered in the design of AI chips, which means processing multiple data points or calculations simultaneously [9].

Custom hardware accelerators are specialized components integrated into AI chips to enhance the performance and efficiency of specific calculations required for artificial intelligence (AI) tasks, especially neural network processing. Previsou study pints out that these accelerators are designed to perform specific operations more effectively than general-purpose processors, thereby helping to improve the overall efficiency of artificial intelligence chips [10]. Customized hardware accelerators can significantly accelerate AI computing speed, enabling AI chips to process data faster and more efficiently. They play a crucial role in promoting the rapid development of artificial intelligence technology in various applications and industries.

Due to its specialized design and intended use, artificial intelligence chips differ significantly from traditional chips in terms of purpose and functionality, architecture and hardware, parallelism and data flow, numerical accuracy, software optimization, and energy efficiency. Firstly, in terms of purpose and functionality, artificial intelligence chips are designed specifically to accelerate artificial intelligence tasks, such as neural network computing for deep learning. They have been optimized for matrix operations, parallel processing, and the computational requirements of processing artificial intelligence algorithms, whether for training or reasoning. Traditional chips aim to perform various tasks across various applications. They execute instructions from various software and are not optimized for specific artificial intelligence workloads. Secondly, in terms of parallelism and data flow, artificial intelligence chips are highly parallel in design and can handle the inherent large-scale parallelism in artificial intelligence computing, especially in neural network processing. They have been optimized for tasks involving simultaneous processing of large matrices and data. Although traditional chips also support parallel processing, their design is more balanced and can handle various tasks, not just artificial intelligence related calculations. Finally, regarding energy efficiency, AI chips are designed specifically for high energy efficiency, aiming to perform AI calculations with minimal power consumption. They excel in providing high performance for specific artificial intelligence tasks while minimizing energy consumption. Traditional chips may provide good performance for a range of tasks, but when performing AI specific calculations, they may not achieve the same level of energy efficiency as AI chips.

4. Configuration and Fabrication

The configuration of AI chips involves setting various parameters and functions in the hardware and software of the chip to optimize its performance and functionality for specific AI workloads. Configuration plays a crucial role in determining the efficiency of chips in processing artificial intelligence tasks, whether it is training complex neural networks or performing real-time inference. Artificial intelligence chips come in various forms to meet different artificial intelligence workloads and applications. The three common types of artificial intelligence chips are GPU (Graphics Processing Unit), ASIC (Application Specific Integrated Circuit), and FPGA (Field Programmable Gate Array). Each type has its own advantages and has been optimized for specific tasks in the field of artificial intelligence. In the 2000s, due to insufficient CPU capabilities in image rendering, GPUs were invented to share this workload and perform polygon image rendering on the screen. The GPU, originally designed for rendering images and graphics in video games, was found to be very effective for parallel processing, making it suitable for artificial intelligence computing. Researchers and developers have started using GPUs to accelerate neural network training, significantly improving speed compared to traditional CPUs. Toru introduced that GPU was used as a general Massively parallel processor in 2006, thanks to the general core of GPU [11]. He also talked about the most concerned autonomous vehicle technology, which also needs the application of GPU. The advantage of GPU lies in its high parallel processing ability and flexibility, but its limitations are also very

obvious. The large energy consumption of GPU can affect its efficiency in certain environments, and GPU has not been optimized for low latency real-time inference tasks.

In 2010, with the rapid growth of artificial intelligence applications, the demand for more professional and efficient hardware increased. This led to the development of specialized integrated circuits (ASICs) specifically designed for artificial intelligence workloads. ASIC can be optimized for specific AI tasks, achieving better performance and energy efficiency compared to GPUs and CPUs [12]. As a specialized AI chip tailored to meet specific application needs, ASIC can significantly improve performance while also driving the specialization of artificial intelligence chips for downstream needs, many of which cannot rely solely on cloud technology, For example, autonomous cars, drones, and various smart appliances require cognitive interaction and artificial intelligence computing capabilities. Due to the need for local hardware infrastructure support, there is also a huge demand for artificial intelligence chips. Throughout the 2010s, the adoption of artificial intelligence chips in various industries continued to increase. From smart phones and Edge device to data centers and cloud infrastructure, AI chips have become an indispensable part of accelerating AI workloads and implementing real-time AI applications. ASICs are custom designed chips built specifically for specific applications or workloads. The advantage lies in its performance and efficiency, and ASIC can be designed to perform specific artificial intelligence operations with excellent speed and energy efficiency. At the same time, ASIC is customized for specific tasks to minimize unnecessary hardware components and reduce power consumption. However, similarly, ASICs also lack flexibility because once designed and manufactured, they are not easily reconfigured for different tasks, making them suitable for specific use cases. Moreover, designing and manufacturing ASICs is a time-consuming and costly process.

FPGA is a reprogrammable hardware chip that can be configured and reconfigured after manufacturing. They allow users to create custom digital circuits. Unlike specialized integrated circuits (ASICs), FPGAs are not designed for specific applications during the manufacturing process, but are programmed by users to perform specific functions. AI chips based on Field Programmable Gate Arrays (FPGAs) provide a unique way to accelerate artificial intelligence (AI) workloads. These FPGA based AI chips provide customizable hardware acceleration for specific AI operations, which can improve performance and energy efficiency compared to traditional processors. Previsou reserach provides a detailed introduction to FPGA [13]. In terms of energy efficiency, compared to traditional CPUs or GPUs, FPGA based AI chips can achieve high performance with lower power consumption. This efficiency is particularly advantageous for edge devices and other scenarios with limited power resources. Meanwhile, FPGA has strong reconfigurability, and one of its outstanding features is the ability to reprogram for different tasks. This enables them to adapt to the constantly evolving workload of artificial intelligence, allowing for rapid updates without the need to change the underlying hardware.

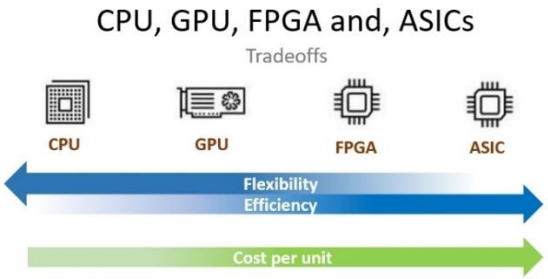

Figure 1. Tradeoffs between different chips

With the advancement of FPGA technology and the emergence of more user-friendly development tools and frameworks, FPGA based AI chips have become more accessible and practical for a wider range of AI developers and applications. They provide convincing choices for achieving high performance, energy efficiency, and customization in artificial intelligence computing, especially in scenarios where real-time processing and edge artificial intelligence are key considerations. A comparison and trade off diagram for different chips is shown in Fig. 1 [14]. In summary, GPU, ASIC,

and FPGA each have their own advantages and disadvantages in AI computing. GPU provides high parallel processing capability and flexibility, ASIC provides optimized performance and efficiency for specific tasks, and FPGA provides a balance between flexibility and customization. The choice of which type of AI chip to use depends on the specific requirements of the AI workload, the trade-off between performance and efficiency, and the available development resources.

5. Applications

AI chips are widely used in various industries due to their ability to accelerate artificial intelligence (AI) computing and efficiently handle complex data processing tasks. These chips play a crucial role in achieving advanced artificial intelligence technology and applications. Some noteworthy applications of AI chips include image and video processing, natural language processing (NLP), deep learning and neural networks, autonomous vehicle, healthcare, finance and trading, manufacturing and industry, robotics, drug discovery and biotechnology, energy management, environmental monitoring, and games and entertainment. With the increasing integration of artificial intelligence technology into our daily lives, the application range of artificial intelligence chips is still expanding. Image and video processing is a prominent application field of artificial intelligence chips, as it involves complex computational tasks and requires effective processing of large amounts of visual data. Artificial intelligence chips play a crucial role in accelerating these tasks, enabling real-time analysis, interpretation, and operation of images and videos. Schoalrs fully illustrates this point [15]. Artificial intelligence chips can be used for image recognition tasks, analyzing images, and classifying objects, scenes, or patterns within them. Not only in healthcare (medical image analysis), but also in industries such as retail (product recognition) and security (monitoring systems).

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) dedicated to enabling computers to understand, interpret, and generate human language in a meaningful and contextual manner. AI chips play a crucial role in accelerating the computational requirements of NLP tasks, which involve processing and analyzing text data. NLP includes various tasks aimed at understanding the meaning of text. Previsou study provides a more detailed explanation of the role of NLP and AI chips, which can accelerate the processing of text data to recognize patterns, emotions, and key entities [16]. NLP also involves generating human like text based on input data. This includes tasks such as text summarization, language translation, chat robot response, and content generation. Artificial intelligence chips help generate coherent and contextual text responses. Artificial intelligence chips also include speech recognition and synthesis, named entity recognition (NER), emotion analysis, language translation, language modeling, real-time interaction, and semantic understanding for NLP. With the continuous growth of NLP applications, AI chips will continue to be a key driver of efficient and accurate language processing. The synergy between the advancement of NLP algorithms and the computing power of artificial intelligence chips is changing the way one interacts, access information, and communicate with machines in the digital age.

Artificial intelligence chips have a significant impact on the field of robotics by achieving advanced functionality, autonomy, and intelligence in robotic systems. These chips enable robots to perceive the environment, make real-time decisions, and efficiently and accurately perform complex tasks. A very detailed description is provided on how AI chips can be applied in the field of robotics [17]. Firstly, in terms of sensing and perception, artificial intelligence chips process data from sensors such as cameras, LiDAR, and ultrasonic sensors to create detailed maps of the robot's surrounding environment. These maps help robots navigate, avoid obstacles, and understand spatial backgrounds. Secondly, artificial intelligence chips can enable robots to detect and recognize objects in the environment. This is crucial for tasks such as picking and placing objects, interacting with the environment, and human-machine collaboration. For operation and grasping, artificial intelligence chips help analyze the shape, size, and texture of objects, enabling robots to accurately grasp objects. They also help to manipulate objects in a clever way. Artificial intelligence chips can improve the accuracy and consistency of tasks that require repetitive operations, such as assembly, manufacturing,

and quality control, in handling the accuracy and repeatability of robots. For some special purpose robots, such as medical robots: artificial intelligence chips process sensor data to help accurately move, analyze patient data, and make real-time adjustments during surgery. Entertainment and service robots: Artificial intelligence chips can provide power for humanoid and social robots used in entertainment, hotel, and customer service environments, enabling interactive dialogue and personalized interaction. Search and rescue robots: Artificial intelligence chips enhance the perception and decision-making abilities of search and rescue robots, helping to find survivors in disaster areas. In short, artificial intelligence chips are at the forefront of advanced robotics technology, enabling robots to operate autonomously, learn from the environment, and adapt to new situations. The synergy between artificial intelligence and robotics technology is driving innovation in automation, safety, efficiency, and human-machine collaboration across various industries.

With the increasing integration of artificial intelligence technology into our daily lives, the application range of artificial intelligence chips continues to expand. These chips are at the forefront of innovation, promoting progress in multiple industries and driving the development of artificial intelligence solutions.

6. Limitations and Prospects

Artificial intelligence chips have changed the landscape of artificial intelligence by enabling faster and more efficient complex computing processing. However, like any technology, artificial intelligence chips also have their limitations and face ongoing challenges. The design focus of many artificial intelligence chips is on specific artificial intelligence workloads, which may limit their versatility in handling various tasks beyond specialization. Although artificial intelligence chips provide higher energy efficiency compared to traditional processors, some highly specialized artificial intelligence chips may optimize performance at the expense of energy consumption. As mentioned earlier, AI chips such as ASICs and FPGAs require customization for specific tasks, which may result in longer development cycles and reduced flexibility to adapt to constantly changing workloads. In addition, as AI chips become increasingly powerful, people's concerns about their potential abuse, bias, and unintended consequences in the decision-making process are also increasing. Even so, the future of AI chips is full of infinite possibilities. With the technological progress and research progress in artificial intelligence, material science, and chip design, artificial intelligence chips will play an important role in shaping the development trajectory of artificial intelligence, bringing us closer to achieving more intelligent, efficient, and adaptable systems.

7. Conclusion

In summary, this study provides a comprehensive overview of the principles and state-of-the-art applications of artificial intelligence chips. This includes the history of artificial intelligence, the development of artificial intelligence chips, and their role in advancing the field of artificial intelligence. From natural language processing to computer vision, artificial intelligence chips have created intelligent systems that can efficiently perform complex tasks. With the continuous development of artificial intelligence technology, one can expect artificial intelligence chips to play an increasingly important role in shaping the future of computing and artificial intelligence research.

References

[1] Turing A M. Computing machinery and intelligence. Springer Netherlands, 2009.

[2] Ruhmelhart D, Hinton G, Wiliams R. Learning representations by back-propagation errors. Nature, 1986, 323 (533-536): 10.

[3] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015, 521: 436–444.

[4] Gupta N. Introduction to hardware accelerator systems for artificial intelligence and machine learning. Advances in Computers. Elsevier, 2021, 122: 1-21.

[5] Momose H, Kaneko T, Asai T. Systems and circuits for AI chips and their trends. Japanese Journal of Applied Physics, 2020, 59 (5): 050502.

[6] Pang G. The AI Chip Race. IEEE Intelligent Systems, 2022, 37 (2): 111-112.

[7] Jouppi N, Young C, Patil N, Patterson D. Motivation for and Evaluation of the First Tensor Processing Unit. IEEE Micro, 2020, 38 (3): 10-19.

[8] Samuel G. Making chips smarter. Communications of the ACM, 2017, 60 (5): 13–15.

[9] Song L, Chen F, Chen Y, Li H. Parallelism in Deep Learning Accelerators. 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, China, 2020: 645-650.

[10] Wang E, Davis J J, Zhao R, et al. Deep neural network approximation for custom hardware: Where we've been, where we're going. ACM Computing Surveys (CSUR), 2019, 52 (2): 1-39.

[11] Baji T. GPU: the biggest key processor for AI and parallel processing. Photomask Japan 2017: XXIV Symposium on Photomask and Next-Generation Lithography Mask Technology. SPIE, 2017, 10454: 24-29.

[12] Li B, Gu J, Jiang W. Artificial intelligence (AI) chip technology review. 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2019: 114-117.

[13] Li Z, Zhang Y, Wang J, et al. A survey of FPGA design for AI era. Journal of Semiconductors, 2020, 41 (2): 021402.

[14] Fecotrons website. Retrieved from: https://Fecotrons.com/Fnews/Fai-computing-chip-analysis-for-software-defined-vehicles-blog%2F&psig=AOvVaw3KOccHfO2KR5_0IC46YMoQ&ust=1693316154919000&source=images&cd=vfe&opi=89978449&ved=2ahUKEwiL0sKwvP-AAxUVu4kEHS7hCq4QjRx6BAgAEAw.

[15] Konnova N S, Basarab M A, Basarab D A. Image processing using artificial intelligence methods in cardiovascular decision support systems. Proc. SPIE International Conference on Image and Video Processing, and Artificial Intelligence, 2018, 108361U (29 October 2018).

[16] Nagarhalli T P, Mhatre S, Patil S, Patil P. The Review of Natural Language Processing Applications with Emphasis on Machine Learning Implementations. 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022: 1353-1358.

[17] Lester A G. Robotics and Artificial Intelligence. Springer Science & Business Media, 2012.