symmetry

MDPI

Review

Artificial Intelligence: An Energy Efficiency Tool for Enhanced High performance computing

Anabi Hilary Kelechi [1], Mohammed H. Alsharif [2], Okpe Jonah Bameyi [1], Paul Joan Ezra [1], Iorshase Kator Joseph [1], Aaron-Anthony Atayero [1], Zong Woo Geem [3,*] and Junhee Hong [3,*]

[1] Department of Electrical Engineering and Information Engineering, College of Engineering, Covenant University, Canaanland, Ota P.M.B 1023, Ogun State 110125, Nigeria; hilary.anabi@covenantuniversity.edu.ng (A.H.K.); okpe.jonah@stu.cu.edu.ng (O.J.B.), joan.paulpgs@stu.cu.edu.ng (P.J.E.), kator.iorshasepgs@stu.cu.edu.ng (I.K.J.), atayero@covenantuniversity.edu.ng (A.A.A.)

[2] Department of Electrical Engineering, College of Electronics and Information Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Korea; malsharif@sejong.ac.kr

[3] Department of Energy IT, Gachon University, Seongnam 13120, Korea

* Correspondence: geem@gachon.ac.kr (Z.W.G.), hongpa@gachon.ac.kr (J.H.)

Received: 16 May 2020; Accepted: 16 June 2020; Published: 18 June 2020

Abstract: Power-consuming entities such as high performance computing (HPC) sites and large data centers are growing with the advance in information technology. In business, HPC is used to enhance the product delivery time, reduce the production cost, and decrease the time it takes to develop a new product. Today's high level of computing power from supercomputers comes at the expense of consuming large amounts of electric power. It is necessary to consider reducing the energy required by the computing systems and the resources needed to operate these computing systems to minimize the energy utilized by HPC entities. The database could improve system energy efficiency by sampling all the components' power consumption at regular intervals and the information contained in a database. The information stored in the database will serve as input data for energy-efficiency optimization. More so, device workload information and different usage metrics are stored in the database. There has been strong momentum in the area of artificial intelligence (AI) as a tool for optimizing and processing automation by leveraging on already existing information. This paper discusses ideas for improving energy efficiency for HPC using AI.

Keywords: 5G; high performance computing (HPC); artificial intelligence (AI); energy efficiency (EE); machine learning (ML); Big Data; Internet of Things (IoT)

1. Introduction

The sheer magnitude of data that companies are currently exposed is on the increase as a result of emerging technologies such as the Internet of Things (IoT), artificial intelligence (AI), data in motion (DIM), and 3-D imaging [1–3]. Real-time data processing is non-negotiable in today's mobile content driven environment for reasons such as live broadcasts of sporting events, monitoring a growing hurricane, checking new products, and evaluating stock trends. Organizations need a highly efficient, lightning-fast information technology (IT) infrastructure to handle, store, and analyze vast quantities of data to stay a step ahead of their competitors [4–6]. AI, IoT, and Fifth Generation (5G) communications would be critical drivers of high performance computing (HPC) development as they allow vast amounts of data processed at a very high speed [7]. An example of the HPC architecture is NVIDIA, which enables new applications to use the same computing network and it is expected that 5G radio access network will adopt the technology as well to meet with its computing requirements [8]. HPC remains an essential tool for scientists today. HPC allows scientific advancement by simulating situations where trials are either impossible or theory only is insufficient [9–11]. The high level of computing power obtainable from recent supercomputers detrimental as it consumes enormous amounts of electrical energy [12,13]. Currently, energy efficiency (EE) is a critical problem for HPC sustainable development and has been addressed from different perspectives such as; network, input, output, and resource organization. The industry has concentrated on built-in and low-power computing infrastructures using reduced instruction set computing (RISC) processors for an increase in EE [14,15]. HPC centers use a massive volume of power to operate powerful computers and infrastructures required to cool [16].

As could be seen in Figure 1 and Figure 2, the microprocessor supply voltage and power have been on the decrease relative to computational enhancement. Although the 2015-to-2020 dataset was not included because it is not available, not much trend reversal is expected because no novel technology has emerged to address this issue. Figure 1 illustrates that the microprocessor voltage has not scaled down considerably, just in the order of 8% per year. Figure 2 indicates that over the past decades microprocessor power consumption has increased exponentially. In both Figures, the year 2020 was not included because the year 2020 is not yet over.

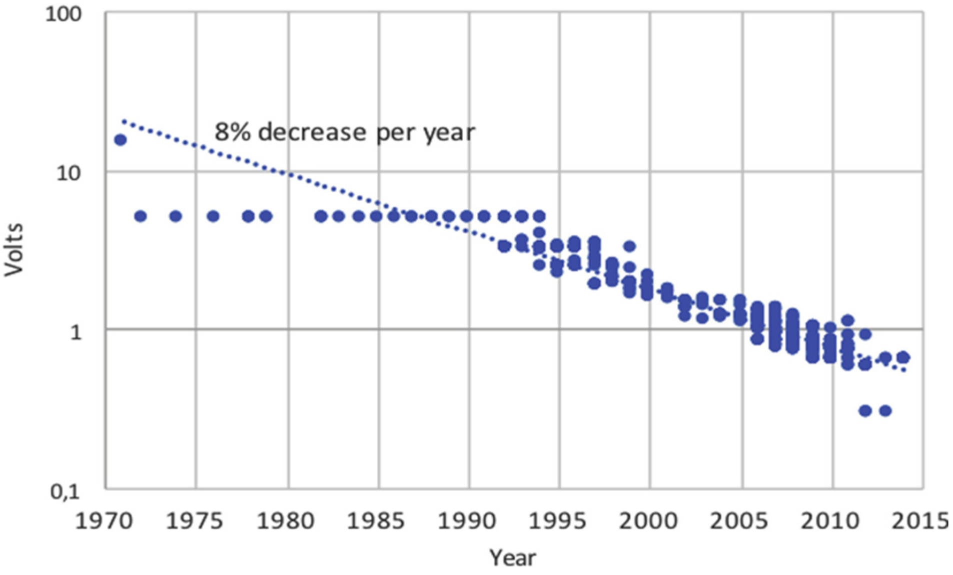

Figure 1. Minimum supply voltage of microprocessors [17].

Recently, significant efficiency improvements have resulted from advancements in multicore as well as accelerator technology. Computing centers quickly embraced this technology to meet the growing request for computing power resources. Electric power consumption is a significant cost element for data centers [18–20].

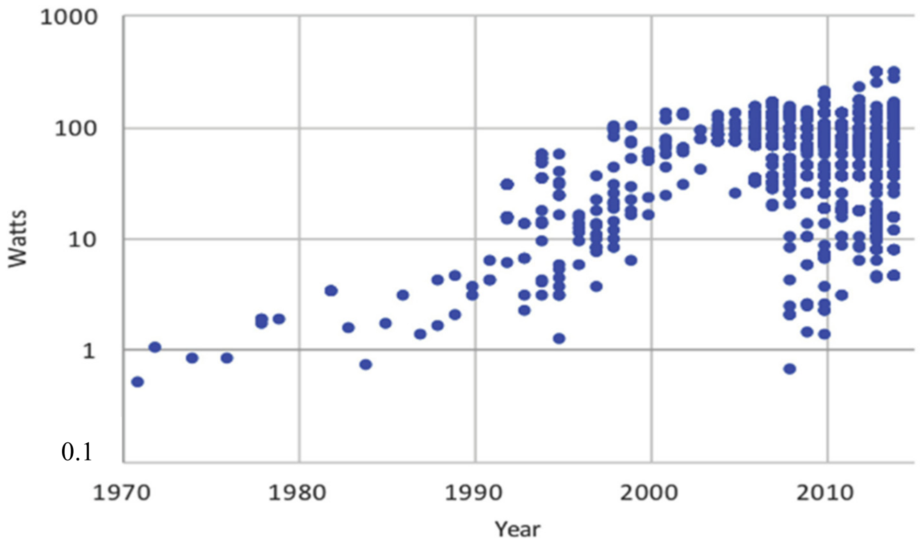

Figure 2. Power (Thermal Design Power, TDP) of microprocessors [17].

Besides, energy consumption increases carbon dioxide emissions, carbon footprints, climate, and human health hazards. Also, the heat energy reduces the efficiency and service life of the hardware module [20,21]. The synergy between HPC and AI can be a thrilling revolutionary model both for business and for technology. AI technological innovations are rapidly growing study fields, motivating big-data analysis. A convergence of AI and HPC could be used to evaluate various software applications without necessarily creating a reference model [22].

The motivation driving the advance in computing is AI and has shaped computing the way we know it today. AI returns to inspire computation in [23]. Though EE has been extensively discussed for ICT before now [12]; however, not many HPC sites have adopted flexible energy-saving methods. HPC sites and large data centers tend to belong to the largest power-consuming organizations in the ICT sector presently [13,24]. Most of those sites run systems with processing cores counting in thousands. With the increase in processor numbers, the power utilization of the systems becomes more significant. Up until now, mobile apps and technical data centers have become the critical drivers of energy-efficient ICT, finding an improved battery running times, and reducing functional costs. Regrettably, several of the energy-saving machineries built either do not measure up to or are not relevant at all to the dimensions of HPC sites [12]. This work aims to discuss substantial issues to address the problem of EE within HPC. To this end, this study attempted to incorporate as many directions as possible. Restricted by size constraints, this work deeply investigated controversial research topics based on their respective sub-domains to achieve a precise, concrete, and concise conclusion. The key contributions of this study are summarized as follows:

This study presents a comprehensive overview of research topics on using AI to address the problem of EE within HPC. This issue was deeply investigated based on their respective sub-domains to achieve a precise, concrete, and concise conclusion.

This study discusses the current and future technologies for improving the energy efficiency of HPC applications and infrastructures.

For researchers, this article will contribute significantly to opening new horizons for future research directions by providing several new references that could support the use of AI to address the problem of EE within HPC.

The remainder of the paper is organized in the following way. Section 2 provides context information on the history of HPC evolution. Section 3 contains an analysis of relevant literature, and Section 4 offers an overview of HPC in 5G Networks and energy efficiency. In Section 5, we examined artificial intelligence and how AI tools address the issues of energy efficiency in HPC. Section 6 presents a practical application and limitations and conclude in Section 7.

2. HPC Overview

HPC evolved due to the increase in the demands for processing speed. HPC brings different technologies such as algorithms, programs, electronics, and system software to solve advanced problems effectively. A highly-efficient HPC system requires a high-bandwidth, low-latency network. In 1960, the high performance computer was introduced into the market by Seymour Cray at Control Data Corporation (CDC). This computer grew more reliable and faster with extra core processors. In this section, we educate the readers on what HPC is, why it is critical in modern communications, and how it works.

2.1. What is HPC

HPC's can process data and undertake complex calculations at high speeds. Take, for instance, a 3 GHz processor in a laptop or desktop can perform 3 billion calculations per second. Ordinarily, this is fast from a human perspective. HPC, on the other hand, performs quadrillions of mathematical computations in a second. The most popular type of HPC is the supercomputer. A typical supercomputer consists of thousands of computing nodes wired parallel working together to complete a task. In 1964, CDC 6600 was the leading supercomputing machine equipped with a single processor that can carry out 3 million calculations per second. Still, the modern smartphone is tens of thousands of times faster [25]. In 1990, HPC released slower processing speeds than an iPhoneX. TeraFLOPS (floating-point operations per second) is the metric for measuring supercomputer's computing capability. Scientists have used HPCs to generate climate models to providing visual insight into climate evolution. They have also taken to skies by storm and launching supercomputers into space for exploration and data collection. The HPC system design and architecture have multifaceted challenges [26,27].

2.2. Why HPC is Important to Modern Communication

Discovering new frontiers involves assembling massive data chunks and making some intelligence out of them. When fully deployed, IoT technology can generate trillions of data samples over some instances of time. These massive data are the vehicle on which data intelligence drives. All the apps on our mobile gadgets are deposited in the cloud, which is a software platform. Software are driven by hardware, and the hardware is located in the data centers. The hallmark of modern communication is the ever-presence of massive data centers whose function is to store and warehouse application context needed by end-users. To fully understand the configuration of the datacenter, the reader is referred to [28–30]. Energy and power management is one critical problem considering the growth phase of an Exascale network. Excessive power consumption is a crucial drawback mitigating against the scaling up of HPC systems [26]. The adaptive power management system is, therefore, vital to the design and operation of HPC systems to boost EE [31]. The scheduling of energy and power-aware jobs and the managing of resources are quite critical for improvement in EE [32,33]. The efficient power monitoring system may include the following attributes; self-awareness, self-directed resolution-making, and the potential to classify jobs online. Others are plug-in regulating algorithms, including time-based power measurement with smooth imaging and effective analytical methods [28,34].

It is a tough job to achieve ExaFLOP output within a 20-megawatt targeted power consumption. Also, the high load variability due to the recurrent change between the computation phases of HPC applications needs different power rates at different periods. It will require the development of intelligent systems, i.e., a system that knows, anticipate by learning, and makes the necessary decisions to ushering an efficient energy management system. It implies that AI would be central to the operations and management of the next-generation HPC Program [16,22,35].

2.3. What are the Operational Modalities of HPC

HPC is a leading field of computer science which concentrates on supercomputer architecture, parallel algorithms, and parallel software development [24]. As a cutting edge technology, supercomputing has always been a specialized type of computing. Moreover, the field of computing expands and evolves, computing has become broader and more complex [25]. A functional HPC, as shown in Figure 1, consists of three (3) key components, namely: i) *Compute nodes*; ii) *Network*; and iii) *Storage*. As could be seen from Figure 3, the compute nodes servers are networked together to form *clusters*. On each of the compute nodes servers, software, and algorithm run simultaneously in the clusters. All the clusters are networked, and their output is kept in the data storage.

The trend in HPC as of 2019, begun with HPC democratization; however, most HPC work continues to be performed in-house, in dedicated or private clouds. The HPC workload in the public cloud continues to expand, and sizeable open house providers such as amazon web services and Microsoft Azure attract household users [1,3,7]. The rise in the graphics processing unit (GPU) also led to massive growth in HPC. From machine learning to self-driving cars, GPU is being used to carry out data-intensive functions. It has proven to be a superior chip for processing and handling HPC workload. Since the growth of GPU computing, AI and HPC have become synonymous with companies like NVIDIA. Google launched its tensor processing unit (TPU). A TPU is an AI accelerator application-specific integrated circuit (ASIC) developed by Google specifically for neural network machine learning. The future of HPCs focuses on efficiency, HPC aggregate data processing power to deliver efficient, reliable, and rapid results. HPC is taking the problem-solving to the next level, doing more with less.

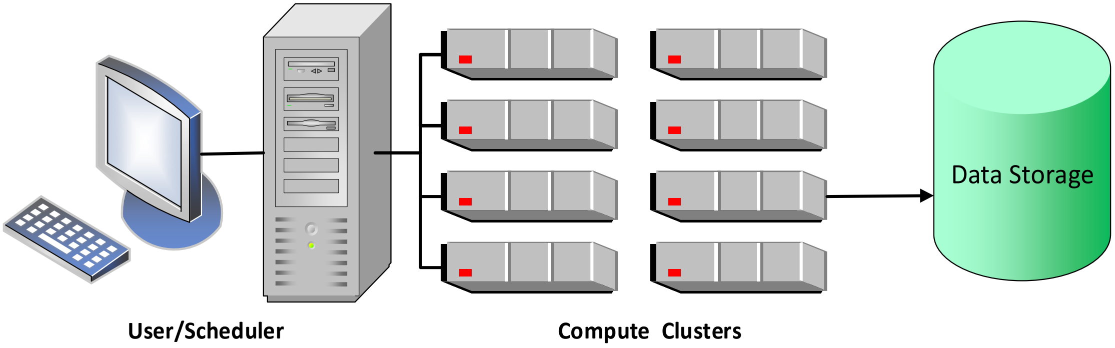

Figure 3. High performance computing (HPC) deployment scenario.

3. Comparative Studies

EE has become one of the most sought-after design parameters for the current computer systems, especially the large-scale structures. Several strategies are presently explored for enhancing EE for HPC systems, both in terms of architectural design, hardware, and software technologies [36–38]. The first study of energy-efficient Ethernet (EEE) in the field of HPC was presented by [39] via assessing its power-savings capacity. In contrast with previous proposals, a thorough study of the effect of added EEE latency overhead was provided using several virtual systems applying traces of real HPC applications. The concept of "power-down threshold" was proposed as a potential addition to EEE to reduce the on/off overhead changeover. The studies discovered that EEE saves approximately 70 percent in connecting the power by shutting off connections, but at the expense of efficiency, leading to a 15 percent (average) increase in total system power consumption. The authors of [40,41] focused their study on the description of some evolutionary changes in HPC hardware, and how recent hardware trends pose challenges associated with Exascale computing hardware development. Reference [42] examined energy management problems, challenges, and potential solutions for the period 2010-2016 by concentrating on the energy usage of data centers and HPC systems. The EE issues currently affecting data centers were highlighted, potential threats identified, as well as several short-term predictions. Additionally, the study grouped energy-efficient approaches into seven components and also Exascale as an HPC framework prospect. Reference [43] described and analyzed several methods of presenting the energy consumption of HPC systems at runtime and a method for estimating the energy ingestion of protocols for fault tolerance. A strategy to categorize fault tolerance protocols into three groups of families; (hierarchical, coordinated, and uncoordinated) was advocated and showed how important the strategy would help users make correct choices concerning energy-efficient services. Reference [44], studied the correlation among both EE and strength of large-scale parallel systems. It was illustrated theoretically and empirically that significant energy savings are possible by merging undervolting and conventional software-level stability methods on contemporary HPC systems without the need for hardware redesign. The system is evaluated experimentally and shown to save up to 12.1 percent energy relative to the reference runs of 8 HPC specifications. Furthermore, it can save up to 9.1 percent more energy than a state-of-the-art frequency regulated dynamic voltage and frequency scaling (DVFS) solution; lower operating frequency or hardware device voltage supply [31,33]. DVFS is a significant way of reducing a computer system's power and energy usage because CMOS-based parts (e.g., CPU, GPU, and memory) are the key power consumers in the device. Reference [45], provided a study of AI-based energy building forecasting procedures with a particular concentration on ensemble models. Four major types of AI-based forecasting have been researched based on concepts and implementations, including multiple linear regression, artificial neural networks (ANNs), supporting vector regression, and set model. This paper also addressed the advantages and disadvantages of each type of model. The paper carried out an intensive discussion of the advantages and disadvantages of AI-based prediction models. Reference [46] stated that diverse researchers describe AI in various ways. There are two dimensions to the differences in the AI definition: One is human centrality, and the other is rationality. Most of the aspects that the intelligence deals with rational actions are adopted. Reference [47], did not consider the use of AI in HPC to make energy efficient. They suggested HPC AI500—a test suite to analyze HPC systems that run scientific workloads on the DL. The growing workload from HPCAI500 is focused on real-world, scientific DL applications, spanning the most representative scientific fields. They proposed a set of metrics for the thorough evaluation of HPC AI systems, taking into account both accuracy, efficiency, power, and cost. Reference [48] informs of a study conducted in machine learning concentrating on refining the predictive performance of algorithms, but recently, researchers are becoming more interested in improving EE as well. The paper gives an insight as to why developing energy-efficient algorithms in machine learning is of great importance. In comparison to past methods, AI enables HPC systems beyond basic rules-based instructions. Reference [49] indicated that in comparison to previous methods, AI empowers HPC systems beyond basic rule-based instructions. Instead, AI tests the data using a series of 'theories' and algorithms as instructions. Reference [50], suggested two initiatives (Machine Learning classifiers and DVFS settings during runtime) to address balancing application performance and system power consumption in HPC during runtime of the program, using closed loop feedback architectures based on the self-aware computing paradigm to observe, decide, and act, presented ultramodern energy-conscious HPC, particularly the recognition and grouping of strategies by device and unit size, optimization metrics, and energy or power management methods. Types of system include single computers, clusters, networks, and clouds, while devices comprise CPUs, GPUs, multiprocessors, and hybrid-systems. With respect to modern HPC systems, they addressed tools and APIs, as well as environments aimed at predicting and simulating energy and power intake. Reference [51] gave an overview of the recent research advancements in energy-efficient computing, identified common characteristics, and classified the approaches. They addressed the causes and issues of high power or energy usage and present a taxonomy of energy-efficient computer system design covering the levels of hardware, operating system, virtualization, and data centers. Reference [52] stated that, HPC systems of significant size, system-wide power consumption has been described as one of the core constraints going forward, where DRAM main memory units account for approximately 30-50 per cent of the overall power utilization of a node. Nonetheless, as an alternative to DRAM, a range of new memory technologies called nonvolatile memory (NVM) products are being examined. Reference [53] examines the trade-off between energy and performance (time of execution) for HPC applications in a real small-scale power-scalable cluster as well as the trade-off between energy and performance (time of execution) for serial and parallel HPC programs. From the array of literature discussed, it could be seen that some works have dealt on deploying AI to enhance the EE of HPC systems. Unfortunately, these works have not failed to provide thorough evidence on why AI is needed in HPC systems. Hence, presenting a myopic view. Secondly, there was no linkage between HPC, 5G, and EE in the previous works. Table 1 presents a summary of reviewed related work.

Table 1. Summary on related studies on HPC energy efficiency.

Reference	Objectives
Axel Auweter and Herbert Huber, 2011, [12]	• The paper provides a summary of the energy-efficient operating concepts of an HPC site and reveals variations and similarities with ordinary data centers.
Czarnul, Proficz, and Krzywaniak, 2019, [18]	• The paper addresses state-of-the-art high-performance energy-aware computing (HPC), explicitly identifying and categorizing system and device type strategies, optimizing metrics, and energy/power management methods. The review established several open spaces and significant up-to-date issues relating to methods and resources for modern HPC systems that enable energy-aware processing.
Diouri et al., 2013, [43]	• The paper discusses two possible strategies (with or without knowledge of software and services) with the same objective: To reduce the energy consumption of large-scale systems supporting HPC software. The paper also highlights the importance of helping consumers make the right decisions about energy-efficient services.
Florez, Pecero, Emeras, and Barrios, 2017, [14]	• The authors used an ARM-built cluster, called a millicluster, designed to provide high energy output at low power. A model was developed for estimating energy consumption founded on experimental findings, derived from measurements carried out during a benchmarking process representative of a real-life workload.
Hussain, Wahid, Shah, Akhunzada, and Arshad, 2018, [42]	• This paper analyzed the problems, challenges, and solutions proposed for the period 2010-2016 by focusing on data center and HPC energy use. They classified existing energy management issues currently faced by data centers.
Jiang et al., 2018, [47]	• They introduced HPC AI500 in this paper—a benchmark suite for testing HPC systems that run scientific DL workloads. The workload from HPCAI500, covering the most representative scientific fields, is focused on implementations of real-world experimental deep learning (DL). A collection of metrics was proposed to evaluate the HPC AI systems comprehensively, taking into account both accuracy and performance.
Johnsson, Ahlin, and Wang, 2010 ,[37]	• The paper introduces a concept, one of many under the Partnership for Advanced Computing in Europe (PRACE) initiative, to investigate energy efficiency improvements. The paper addressed a study on system design and preliminary performance outcomes, concentrating on the energy dimensions of tests and comparing findings with Blue Gene/P.
Labasan, 2016, [54]	• A survey of current work on energy-efficient and power-constrained computing techniques. • The paper addressed an overview of these methods as they refer to a particular case for use in HPC.
Lu, 2017, [23]	• The author discussed why and how AI would continue to inspire and reinvent computation when Moore's law runs out of steam.
Saravanan, Carpenter, and Ramirez, 2013, [39]	• The authors provided the first study of energy efficient Ethernet in the HPC domain, exploring its potential for power savings. They suggested the use of "Power-Down Threshold" as a theoretical extension to the EEE to reduce the on/off overhead transition.

Tan et al., 2015, [44]	• Authors present an undervolting energy-saving strategy that leverages conventional resilience strategies to accommodate increased undervolting failures. The policy is driven by analytical models that capture undervolting impacts and the interplay between energy efficiency and resilience. • Experimental results showed that their method could save up to 12.1% of energy relative to the baseline and save up to 9.1% more energy than a state-of-the-art DVFS solution.
Wang, Zeyu and M.E.Rinker, 2015, [45]	• In particular, the paper presents a thorough analysis of AI-based energy prediction approaches, multiple linear regression, ANNs and support vector regression. • The paper also focused on predictive ensemble models used to forecast energy building. Ensemble models boost the accuracy of predictions by combining multiple predictions.
Wlotzka et al., 2017, [55]	• This paper discusses critical issues of high-performance energy-aware computing. The authors outlined several computational methods commonly used in scientific applications and provided an energy profiling and tracing technique appropriate for the study of device power consumption. • They also addressed energy-saving opportunities in computing using two examples. First, for the conjugate gradient process, energy-aware runtime on shared-memory multicore platforms. Secondly, energy-efficient techniques on the distributed memory clusters for multigrid methods.
Yi and Loia, 2019, [22]	• Paper presented a summary of emerging technologies and suggested recommendations for the implementation of HPC and AI solutions. • It covers clean applications and studies within advanced as well as evolving HPC framework and AI applications scopes.
Graham, Susan L. Snir, Marc Patterson, Cynthia A. 2005, [56]	• A report that presented recommendations after analyzing the state of U.S. supercomputing capacities and related research and development.
X. Mei, Q. Wang, and X. Chu, 2017. [36]	• This paper aimed at exploring the impact of graphics processing unit dynamic voltage and frequency scaling (GPU DVFS) on the application performance and power consumption, and furthermore, on energy conservation.
E. Y. Y. Kan, W. K. Chan, and T. H. Tse 2012., [38]	• The study presents an innovative framework, known as EClass, for general-purpose DVFS processors by recognizing short and repetitive utilization patterns efficiently using machine learning. The algorithm is lightweight and can save up to 52.9% of the energy consumption compared with the classical PAST algorithm.
PRACE,2013, [10].	• How supercomputing drives economic growth; drugs, finance, and climate case studies
A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, 2011.[51]	• Discussed causes and problems of high power/energy consumption and present a taxonomy of energy-efficient design of computing systems covering the hardware, operating system, virtualization, and data center levels.
H. Rong, H. Zhang, S. Xiao, C. Li, and C. Hu, 2016, [24]	• The paper reviews the progress of energy-saving technologies in HPC, energy conservation technologies for computer rooms and renewable energy applications during the construction and operation of data centers.
C. Imes, S. Hofmeyr, and H. Hofmann, 2017, [50]	• Proposed two projects to address balancing application performance and system power consumption in HPC during application runtime, using closed-loop feedback designs based on the self-aware computing model to observe, decide, and act.

J. S. Vetter and S. Mittal, 2015, [52]	• A key contributing factor to system power consumption is a system's main memory. • However, a number of emerging memory technologies—nonvolatile memory (NVM) devices—are being investigated as an alternative for DRAM. Moving forward, these NVM devices may offer several solutions for HPC architectures.
V. W. Freeh et al., 2007, [53]	• This paper analyzes the energy-time trade-off of a wide range of applications—serial and parallel—on a power-scalable cluster, a cluster of frequency and voltage-scalable AMD-64 nodes, each equipped with a power meter was used. They also investigated metrics that can, at runtime, predict when each type of bottleneck occurs.
A. I. Dounis, 2010, [46]	• Paper briefly presents expert systems and computational intelligence (CI) techniques and outlines how they operate. The major objective of this chapter is to illustrate how intelligent agents (IAs), and multi-agent systems (MASs) may play an essential role in conserving energy in buildings.
Intel Corporation, [49]	• Offer practical considerations for HPC managers to incorporate AI into their HPC environment and scale those capabilities to accommodate emerging workloads and increasing end-user demand.
E. Garcia, 2017, [48]	• This position paper argues for the reasons why developing energy-efficient machine learning algorithms is of great importance.

4. The Need for HPC Energy Efficiency in the Evolving 5G Networks

With the standardization of 5G communications over and deployment commenced globally, the volume of generated telecommunications traffic will increase exponentially. A surge in data traffic will be expected because 5G networks will require more data centers, edge computing devices, and IT infrastructures to sustain the expected quality of service (QoS). HPCs are an integral part of data centers, edge computing devices, and IT infrastructures. Consequently, there will be an increase in energy power consumption. Take, for instance, in 2018 before the launch of 5G networks; data centers accounted for about 205 terawatt-hours of electricity usage, which is roughly 1% of all electricity consumption worldwide. The 205 terawatt-hours represent a 6% increase in total power consumption since 2010. This number will still increase when 5G is fully deployed and activated. The expected rise in data centers is because 5G will transform the societies enabling features and capabilities hitherto considered as a mirage. The vision of 5G is to support various user case scenarios applications namely [57,58]; i) *massive machine type communication* (MMTC) is driven by a smart meter, smart agriculture, fleet management; ii) *enhanced mobile broadband* (eMBB) focusing on broadcasting, mobile/wireless/fixed devices, non-sim devices, 4K/8K UHD, virtual reality/ augmented reality; and iii) *critical machine type communication* (CMTC) which comprises of traffic safety and control, remote manufacturing, remote training, industrial applications, and monitoring. Both CMTC and MMTC adopt small packets transmission architecture with negligible metadata (control information).

The MMTC is designed to implement low-cost, low energy solutions, small data volumes in massive numbers while CMTC is responsible for the provision of ultra-reliable low latency communications. These different 5G user case scenarios will aggravate enormous pressure on data centers that house HPC systems. The network components must be able to accommodate the high-speed data transfer between computing servers and data storage. In general, 5G potentials as it relates to HPC can be summarized as:

5G is expected to connect people, things, data, applications, transport systems, and cities in smart networked communication environments [59,60].

It should transport a vast amount of data much faster, reliably connect a massive number of devices, and process very high volumes of data with minimal delay [59,61].

It will change how developers build distributed software systems. Today, design choices are constrained by bandwidth, latency, and cost considerations, but these barriers will fall away rapidly [62].

4.1. Discovering IoT Networks Needs in HPC

The vision of the Internet of Things (IoT) is to provide ubiquitous wireless connectivity to anything whose utility can be enhanced by being connected [63]. This implies that sensors, machines, and devices will be connected in a unique fashion that would improve productivity and efficiency in the industry and impact positively on the overall quality of life. Discovering these new frontiers both from the aspect of network architectures and traffic management will present significant technical challenges [64]. The massive numbers of distributed systems in the IoT will lead to new features in 5G. Ideally, IoT devices are not bulk data generators often on the range of 10-20 data payload. However, with their intense numbers, big volume of data is subsequently generated and must be processed. To carter for the ensuing large chunks of data, organizations wills require highly reliable, high-speed IT infrastructure to process, store, and analyze large amounts of payload and metadata The big data generated by these connected devices will activate the use of AI, which helps to automate analytics. With each successive advancement, computing requirements have increased. As a result, data-driven organizations of all types and sizes are discovering that they need HPC platforms [65]. For even modest 5G applications, developers will need to embrace HPC and cloud-scale tools such as distributed file systems, key-value stores, in-memory data grids, and faster Spark-powered analytics to boost the performance of application services [65].

While current mobile networks can provide some connectivity for new services like automated cars and drones, both 5G and HPC will be essential to unlocking these next-generation services. 5G networks owe their speed to the use of millimeter waves (radio signals between 30 GHz and 300 GHz). These high-frequency waves carry more information than their 4G counterparts (4G operates between 1Ghz and 5Ghz), but they have a shorter range. The implication is that 5G systems require considerably smaller cells compared with the 4G systems. A single transmission tower might service a 4G service area, but a 5G network covering the same area may need 100 or more small, low-cost antennas affixed to streetlights and telephone poles. Telcos will need to deploy HPC capabilities to capture and analyze the vast amounts of data coming from a more significant number of access points and 5G devices. For example, autonomous vehicles and drones relying on 5G services will stream telemetry to multiple 5G antennas and rely on HPC, storage, and AI-powered predictive services to fuse data in near-real-time so that vehicles can operate safely and avoid collisions. It is then apparent that 5G will rely on HPC [8,66,67].

4.2. Energy Efficiency

For quite a few decades, the computing world was governed by an energy-efficiency architecture. It must have been the dynamic force underlying desktop and mobile computing, which converted energy efficiency into improved battery life in mobile computing and reduced monthly energy bills for desktops. Energy efficiency is of particular interest to computational areas with capped energy supply. In an HPC environment, reducing the use of energy per job is essential to optimize the machine's work throughput under power limitations. Energy-efficient computing's main aim is to minimize energy consumption without incurring any noticeable performance slowdowns at the same time. The slowing is also accepted because it decreases energy costs. Much work is devoted to the study of the relationship between energy and efficiency in the HPC domain [54].

4.3. HPC and Energy Efficiency

Reference [12] submitted a preliminary step towards improving energy efficiency in HPC. The work constitutes a clean evaluation of the power usage of the entire HPC system, comprising not only computer nodes, interconnecting networks, and storage devices but also elements of the site facilities for cooling, monitoring, and control. The infrastructure of the supercomputing site comprises of the whole area surrounding the HPC facility, that is, the house, the required power supply equipment, power delivery, and cooling. In general, there are ways to boost energy efficiency at the level of HPC site infrastructures; these include [12]: i) Reduction of electrical losses in wires during transformation; ii) advanced cooling technologies; and (3) waste heat reuse. Reference [14], proposed that energy consumption could be estimated using an empirical energy model, calculating each processor's energy consumption at periodic intervals. To predict any application of the broad range of energy-consuming HPC applications, you can build models using the decision trees method; it automatically picks the best suitable model for the running workload. In modeling energy consumption, linear regression is the model of choice, and to acquire multi-component metric models, multivariate linear regression has been used [28,68]. Modeling of node-level energy consumption, based on usage of main node components measurements, does not consider external causes (e.g., cooling) into account and relies on assigning tasks to the devices. Thus, the energy model is derived from the estimation of resource usage and the details of the job executed.

Data collection, preprocessing, division of dataset, forecasting model selection; were methodologies deployed to obtained the models. Reference [14], envisaged that future research directions are; the adaptation of the current energy model to an algorithm for job planning, maximizing energy usage, analyzing several performance metrics concurrently, and comparing energy efficiency with other HPC energy management technology systems in ARM-based mili-cluster job schedules. References [14,18], indicated that the energy and power management tools available could be classified into two domains: tracking and managing. Subject to the method or provider, a few devices only permit the energy and power consumption to be read. In contrast, others can enable energy and power consumption to be read and restricted. They also stated that some devices are meant to limit energy and power usage only, but indirectly where a consumer can adjust the energy consumption, e.g., system frequency, to minimize consumption. There are, ultimately, other inspired methods that package the above-mentioned low-level drivers in a more accessible way. They further provide information on power monitoring and control as elaborated:

Power Monitoring: The researchers began observing the energy and power usage of the whole system using exterior meters like Watts Up Pro after HPC started concentrating not just on the time of execution of jobs but also on energy efficiency. The primary benefit of such an approach is that it monitors the use of real energy and power consumed. However, on the downside, these external meters cannot reveal the consumption of energy and power of device subcomponents (e.g., CPU, GPU, and memory) [18,69].

Power Controlling: There are several indirect tools or methods which permit control of power consumption and energy. DVFS, also known separately as DFS and DVS, is one method that allows one to control the voltage or frequency of the processor not just to minimize energy or power consumption but also to increase the output at the same time. DVFS is obtainable for both CPUs and GPUs [18,36].

Power Monitoring and Controlling: Most hardware manufacturers have introduced complete power management, which includes energy tracking, power usage, as well as regulating power limits [18].

Energy-saving trading results dominated HPC power-aware research. DVFS approaches have been leveraged to reduce depredating performance by addressing various application characteristics, including communication barriers, I/O delays, load imbalances, or repetitive behaviors [36,70]. Although energy savings are needed to achieve power targets, they will not be enough, as the all-embracing objective of Exascale is to control power usage rather than energy use. Initially, it proposed offline methods, which served as the foundation for assessing potential online methods. Online methods are complicated, as decision-making requires detailed models and predictions of the effect on application phases at different CPU frequencies without preceding application knowledge. [54], presented an overview of past and current research projects on energy-efficient HPC techniques. They addressed the various measuring standards for classifying applications as compute-bound, memory-bound, or I/O-bound, finding appropriate resource-saving candidates.

5. Artificial Intelligence: Overview

AI is a statistical/probabilistic tool that offers the machine the ability to learn via machine learning (ML) algorithms. ML is a collection of tools through which AI problems can be solved. This distinction between AI and ML is essential because rapid developments in ML have contributed a lot to speedy worldwide attention in AI. ML algorithms can be classified as a supervised learning algorithm or unsupervised learning algorithm [71]. In supervised learning, the output is expected to learn from the training data. In other words, given a set of information that the machine has seen before, it should be able to recognize and make an informed decision based on it. While in unsupervised learning, you have an input with no corresponding output.

AI can be broadly divided into; regression problem and classification problem, as depicted in Figure 4a and Figure 4b, respectively. The goal of the regression problem is given an input data vector *x* determine the targeted output label *f(x).* In other words, the regression problem is a mapping function that maps *x* to the domain of *f(x)* [72]. The regression problem maps an input vector to the output target by deploying some algorithm to normalize the data. Normalization enables elements to lie between 0 and 1, thus bringing all the values of numeric columns in the dataset to a standard scale. In this context, the maximum value of *f(x)* is 1. Normalization is an excellent technique to use when there is uncertainty in data distribution or when the distribution is not Gaussian (a bell curve). Precisely, normalization crunches the output data to a given range; thus, making prediction easy. The data crunching tools are usually nonlinear algorithms that convert linear input to a nonlinear output. Regression algorithms suffer from the issue of regularization and over-fitting issues. Some of the notable regression algorithms are linear regression, neural network, ANNs, deep learning neural network [72].

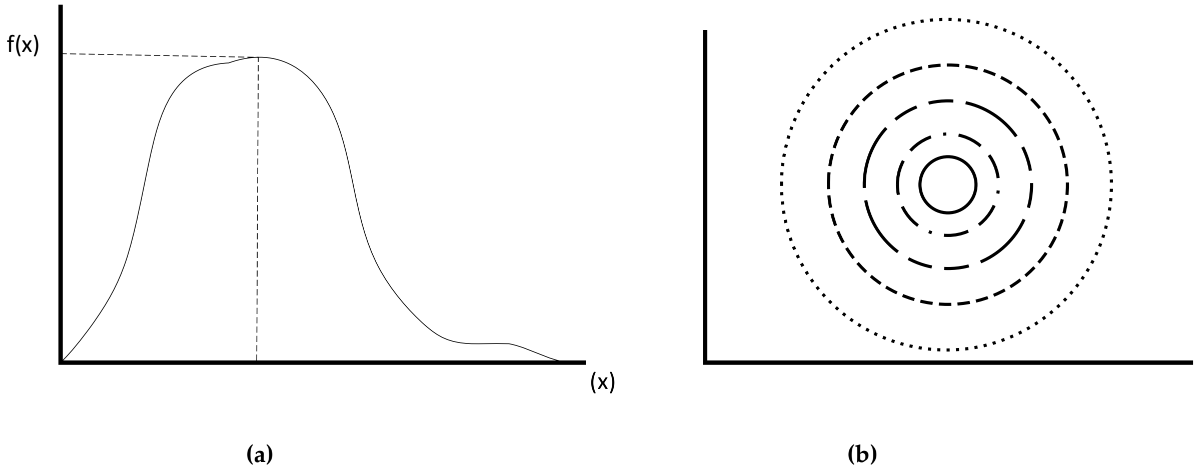

Figure 4. (**a**). Regression task. (**b**). Classification task.

Similarly, classification tends to solve the problem of giving an unknown data set, find the internal morphology which aligns it to a specific pattern. Having acquired the data recognition attribute, one can classify the given object. Hence, it is a pattern recognition task or feature extraction [71]. The goal of the classification algorithm reduces to placing an input vector into one of the centric circles in Figure 4b. Classification algorithms must devise a strategy to overcome outliers. The optimal goal of both regression problem and classification relates to finding that algorithm with the barest cost minimization function. In the context of the regression task, the cost function denotes using that algorithm whose output is as close as possible to the target label. Minimizing cost function can be achieved with the utilization of a gradient descent technique [73]. While in the classification task, the cost minimization problem equates locating that algorithm that clearly understands the internal morphology of the data. Thus, reducing the probability of having outliers. The conventional algorithms used for classification tasks are k-nearest neighbors (KNN), naive Bayes classifier, support vector machine (SVM), classification and regression trees (CART), and random forest. Table 2 notable attributes, as well as advantages and limitations of classification and regression algorithms.

Table 2. Notable attributes, as well as advantages and limitations of classification and regression algorithms.

	Advantages	Disadvantages
Regression Algorithm	Model development is rapid and straightforward. Useful when the relationship to be modeled is not extremely complex and do not have a lot of data.	Applicable only if the solution is linear. In many real-life scenarios, it may not be the case. The algorithm assumes the input residuals (error) to be normally distributed but may not always be satisfied.
Classification	Straight forward implementation. New data can be added seamlessly. Robust against noisy training data. It has the capability to modeling complex classification problems by using many hidden neurons. Maintain the information that presents in the training data.	Does not work well with large dataset except using deep neural network. Sensitive to unbalanced training data. It is a supervised lazy learner. Requires huge memory usage cost

5.1. The Need for AI in HPC

HPC is an aggregation of many compute nodes in clusters that pull their computing resources together, such as computing power to perform a task that would not have been obtained by a single desktop/ workstation. These resources are not infinite as they are faced by energy and cost budgets. Energy and power are two impediments that have slowed down the vast deployment of HPCs as they often capped. These constraints are adopted to minimize operational costs, enhance system efficiency, and increase profitability. Generally, work task in wireless communication is modeled as Poisson Point Process (PPP), which implies work task is a function of some randomness. Hence, the knowledge of work task application potential energy and power consumption will motivate power-cost optimization by scheduling low priority jobs with higher energy/power consumption rates to off-peak hours when the cost of electrical power is cheaper. This knowledge will reduce the overall cost of production and promote a new paradigm in energy costing framework by migrating towards energy-driven charging policies as an alternative to currently existing CPU-hour based charging systems. Historical power/energy data are available and can further be utilized to predict expected needed resources in HPC environment using AI technologies, thus boosting energy efficiency and saving on costs [46,74].

Technological and organizational interventions can affect energy efficiency. Technology programs concentrate on professional development through advanced technology (e.g., new machines or manufacturing procedures). This methodology centers on managing multiple production line schedules to maximize energy usage from distributed sources of energy, like heat and pressure, by leveling or mixing demand and evading significant fluctuations in demand. This approach has been shown to be a success. Organizational procedures can be utilized to boost energy efficiency, especially on short-term performance planning via. energy-oriented planning on a real-time basis using data mining. Recent research intends to further increase energy efficiency by maximizing the energy usage of the entire operation. This technique is being taken by business organizations, which use in-house built technologies to apply AI concepts in their services. That needs a large amount of processing of the data. Google (server centers) and Amazon are among those companies that use AI as an energy-saving tool [75,76].

5.2. AI Tools and Techniques

We have established the fact that AI has found application in ensuring the running of energy-efficient systems. It is utilized to track and optimize energy efficiency, ensure precision in predictions, and optimize power consumption in energy-consuming entities [77,78]. AI imitates the human way of reasoning, learning, and perception to solve complex problems [79]. There are several AI tools and techniques that could be used in HPC to address energy efficiency issues. These include the following; ANNs, multi-agent systems, and reinforcement learning. AI has been tipped earlier on to be crucial in the subsequent phases of HPC System operations and management.

5.2.1. Artificial Neural Networks

ANN is the most easily understood and most commonly used AI model. ANNs learn from training cases and capture interactions among data. ANNs are one of the most representative methods in machine learning because of their robust adaptive learning and generalization capability, especially for nonlinear and non-stationary processes [80,81]. ANNs need few preliminary assumptions to learn from examples by adjusting the relationship's weights [82]. Supervised and unsupervised learning exists. Supervised learning provides the correct output to the ANN for each input sequence. The weights differ to reduce inaccuracy between ANNs input and output specified Exascale [79]. Unsupervised learning provides different patterns of input to the ANN. ANN takes advantage of the relationships between models and discovers how to classify input [72,83]. Some ANNs are a combination of supervised and unsupervised learning. ANN is composed of a node network, arranged in strata. Input nodes obtain input data ln and deliver outputs either by weight $(ln \times w)$ or by decision $(0,1)$ based on a law. Outputs are transferred to the hidden layers consisting of the system's mathematical models. Parameters in the mathematical model involve weights and biases (bn), and there is also communication between hidden layers and output nodes, serving as the output for the outcomes of the AI process. The mathematical model is "trained" over a range of inputs, using outcomes from several established systems. Input from known cases is fed and known outcomes are compared with output [84]. Then the AI program modifies the mathematical model’s weights and biases to provide the model that is a best fit based on the data created.

Training ANNs in AI may be deployed in energy consumption predicting; anomaly detection (comparing input and output data for expected data to identify irregularities once the near fit performance has been established); and energy reduction (Process adjustments can be appraised to accomplish negligible energy assessment). This brings an improvement in energy efficiency for HPCs. ANNs' data structure and nonlinear computations permit good fits to complex, multivariable data [83]. ANNs process data in parallel and are irrepressible to data inaccuracies and can generalize and seek similarities in defective data as long as there are not many neurons in them to overfit data deficiencies. ANN is a complex uninformative model and thus inadequate for process explanatory problems; this is a drawback. When an ANN does not converge, there is no way to say why [83].

5.2.2. Multi-Agent Systems

Multi-agent system (MAS) consists of a network of agents that communicate to meet goals. It is said that an agent is a module of software that contains code and data. It is unable to solve problem on its own the problem assigned to the MAS. The agents interact with each other through a high-level agent communication language (ACL) by exchange of information, request services and negotiate among themselves. Knowledge Query and Manipulation Language (KQML), is the most widely used ACL. It has a layer of communication that covers parameters at low-level such as sender, receiver and identifiers for discussion. A messaging layer determines the performative and interpretive protocol, and a content layer provides another performative information. MASs can model intricate systems with numerous dealings between the autonomous and dynamic units [85]. The efficacy of these depends primarily on the structure of the agent. Network management can be troublesome with peer-to-peer network, whenever a new component is connected to a network, as all agents are updated. For centrally organized infrastructures, only the directory of the facilitator has to be revised with fresh additions. Nevertheless, processing holdups may surface. Other problems in building MASs stem out of their complex form and complex connections among agents, where objectives or distribution of roles and resources can clash [72].

5.2.3. Reinforced Learning

Reinforcement learning (RL) is teaching models of machine learning to create a sequence of decision making. In an unknown, potentially complex environment, the agent learns how to reach a goal [86,87]. An AI in reinforcement learning faces a game-like situation [5]. RL is learning via a learning agent's interaction with its surroundings [88]. The agent learns through trial and error to attain a target. There are three parts to an RL problem namely: Reinforcement function, environment, and value function. This system is complex and probably has a number of states. The atmosphere is dynamic and characterized with a set of states that are possible. For each state *st* at time *t* there is a set A*(st)* of possible actions. The AI gets either bonuses or punishments for the acts it performs to get the system to do what the programmer wants [89]. The aim is to maximize overall reward. Since the designer sets the reward policy, that is, the rules of the game, he does not give the model any clues or suggestions on how to solve the game. It is now left to the model to figure out how to do the job of optimizing the reward, beginning with completely random trials and finishing with advanced techniques and superhuman expertise. RL has been largely restricted to computer programming, and various software and machines use it to determine the best action or direction it should take in a specific situation. Most RL implementations were both in robotics and game play, where RL generates new behaviors instead of modeling existing behaviors. Nevertheless, its use has increased with other AI techniques. Computer programs are increasingly using RL to boost efficiency and output. Energy and Power-Conscious (Aware) Job Scheduling and good Resource Management are crucial in enhancing energy efficiency for this reason, Reinforcement learning is probably the most persuasive way to hint the ingenuity of computer by using the power of search and many tests. As such, for effective power management, control algorithms based on RL are essential algorithms for effective management of energy [72]. The explanation and justification for this is that an adaptive power management architecture is important for the design and operation of HPC systems to increase energy efficiency. Simply because they make up the critical facets of a power management system that is efficient: self-aware, self-governing, and decision-making. Furthermore, they have the capacity to classify workloads in online environment, provide plug-in control algorithms, provide precise spatial and time-based power measurement with smooth imaging and effective analytical techniques.

6. Case Study: A Practical Application

Usage of AI for enhancing energy efficiency can better be comprehended by in view of a real-world application at the Google data centers. Cooling is one key energy consumption source in the data center. As large money-making and industrial facilities, data centers gulp a great deal of energy for cooling, even though a lot of effort has been put in to curtail the growth of energy use, there is still much to do due to the world's increasing need for computing power. Power usage effectiveness (PUE) in the data center setting is a factor by which efficiency is calculated, it is described as the ratio of total energy used up in the building to energy use by IT devices [16]. PUE is a ratio; it is not a reflection of actual consumption of power. With the improvement in the performance of the data center, the overall rate of PUE reduction is still slowed down due to declining yields and the shortcomings of the prevailing cooling technology [84]. Figure 5 shows Google's past PUE presentation from a yearly fleet-wide PUE from 2008 to 2013 [90]. ML is fit and appropriate for the data center setting because of the nature of the plant processes and the numerous monitoring data available. The present, comprehensive data center, features an extensive range of electrical and mechanical equipment, together alongside their set points and controls. It is quite challenging to predict the performance of the data center using conventional engineering formulae; this is a consequence of the exchanges among these systems and different feedback loops.

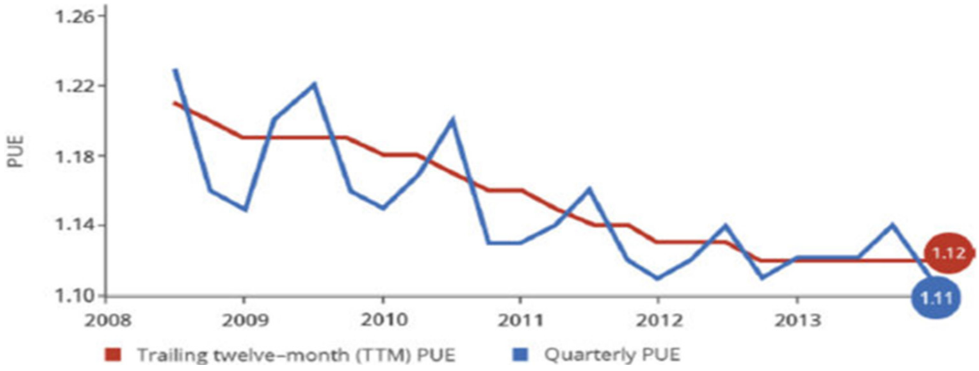

Figure 5. Google's historical power usage effectiveness (PUE) performance [91].

For instance, a simple adjustment to the temperature set point of the cold aisle will result in load differences in the cooling system, resulting in nonlinear changes in equipment efficiency in turn. Since such complex interdependencies are not identified, the use of traditional methods for prognostic modeling most times results in significant mistakes. A large number of potential combinations of equipment and their set point values make it hard to decide where the optimum performance lies [91]. The target setting points can be accomplished from many possible hardware combinations, which may be electrical or mechanical equipment and software, which involves control approaches and set points in a real-time data center. Due to time constraints, regular variations in IT load, and weather conditions, it would be impossible to test every combination of features to optimize performance. Google deployed a machine learning algorithm founded on Deepmind's neural network architecture to solve the problem [91], it is designed to learn from real operational data to mimic and predict plant performance and PUE for a PUE of 1.1 within a 0.4 percent error range respectively. This was achieved by taking historical data; e.g., temperatures, power, pump speeds, and set points already collected inside the data center by thousands of sensors, then use these historical data in the training of a group of neural networks. The goal was to boost energy efficiency at its data center. On average future PUE the neural networks were all trained [72,84]. Two additional deep neural network classes were trained to forecast the impending temperature and pressure of the data center period covering the following hour. These forecasts aimed at mimicking the PUE model's suggested behavior to ensure that operational limits have not been surpassed [91]. Figure 6 is a superior test day outcome, inclusive of periods machine-learning suggestions are switched on or when put off. The machine learning program did achieve a robust reduction of 40 per cent in the total energy used up for cooling [91,92].

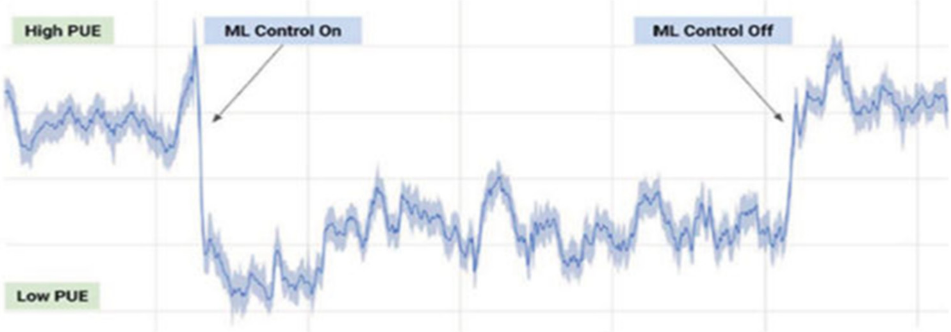

Figure 6. Results of testing [92].

6.1. Model Implementation

The three-layered generic neural network used is shown in Figure 7. The input matrix x for this model is an array $(m \times n)$. The value for training instances is "m". Value for features is "n" (i.e. Data Center input variables), which includes information technology load, weather situations, the quantity of chillers and running cooling towers, setting equipment points. The input matrix x produces the product of the hidden state matrix a; and the model parameters matrix $\theta 1$. In calculating $h\theta(x)$, the output; layer "a" functions as an intermediate condition, the second parameter matrix $\theta 2$ interacts with it. $h\theta(x)$; denotes output variable of attention. It can describe several optimizable standards of measurements [90]. PUE has been chosen in this situation. Understanding the mathematics underlying $h\theta(x)$ action allows the output to be managed and optimized. The last neural network utilized five layers concealed by 50 nodes/hidden layers. Data set for training includes standardized variables for input and at output one standardized variable (data center PUE), each with a resolution of 5 minutes covering 184435-time samples (approximately a 2-year functional data). A large chunk of the dataset is deployed for research (about 70 percent). While the remainder used for the testing and cross validation [72].

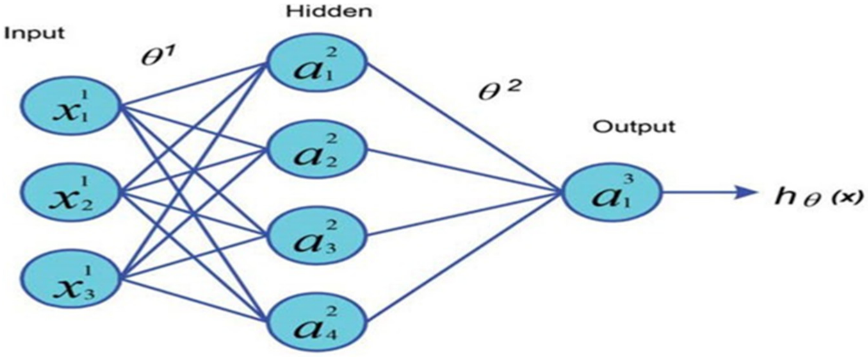

Figure 7. Tree layered neural network [90].

6.2. Benefits of Power Usage Effectiveness

Having a precise and robust PUE model provides plant operators and owners with many benefits. Such benefits include the following:

(1) Automatic performance warning, plant efficiency projections in real-time and troubleshooting using a contrast of actual vs. expected results under a set of conditions stated.

(2) It helps operators of data centers to measure PUE sensitivity to operating parameters of the data center.

(3) It helps operators to model Data Center operational conditions making no physical adjustments or changes. This approach emphasizing simulation enables operators to virtualize the Data Center and describe ideal plant configurations while minimizing the doubt concerning changes in plants.

6.3. Limitations

The value and amount of input data are restricted in machine learning applications, and hence, a limitation. To train the mathematical model accurately, there is a need for a full range of operating situations. With conditions where fewer data exist, the model accuracy may decrease. The same predictive precision can be obtained with many model parameters as for all empirical curve fits. It is at the discretion of the researcher and the operator to apply rational assessment in assessing model forecasts [72,84].

7. Conclusion

This paper gave a common clue of how HPC systems can boost energy efficiency. Specific knowledge bases have been checked to detail ways to run energy-efficient HPCs. It has suggested the fundamental concepts of AI, then suggested appropriate methods of application in enhancing energy efficiency in HPCs. It mentioned AI and HPC and the remarkable hope for the future ahead. We assume that a comprehensive implementation of AI software would make it possible to reduce the power consumed as much as possible and thus increase the production of improved energy-efficient

HPC systems. New experiments will produce results that we still need to have to imagine. HPC transitions in any sector are to be aligned with AI as the new industrial revolution. This paper also reckons that, though current mobile networks can provide some connectivity for new technologies such as autonomous cars and drones, both 5G and HPC must enable these next-generation technologies. It is apparent then that 5G would also depend on HPC.

Author Contributions: O.J.B., P.J.E., and I.K.J. wrote the main parts and conceptualization the first draft of this paper; methodology, A.A.A.; project administration M.H.A. and A.H.K.; A.H.K. and M.H.A. revised the final version of paper. Writing-Review & Editing, Z.W.G., J.H.; Supervision, Z.W.G.; Funding Acquisition, Z.W.G., J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20172010000190).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gill, S.S.; Tuli, S.; Xu, M.; Singh, I.; Singh, K.V.; Lindsay, D.; Tuli, S.; Smirnova, D.; Singh, M.; Jain, U.; et al. Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges. *Internet Things* **2019**, *8*, 100118
2. Tapscott, D.; Tapscott, A. *Blockchain Revolution: How the Technology behind Bitcoin is Changing Money, Business, and the World*; Penguin Publishing Group: London, United Kingdom, 2016.
3. Martin, C.; Leurent, H. Technology and innovation for the future of production: Accelerating value creation. In World Economic Forum: Geneva, Switzerland; 2017.
4. NetApp. What is High Performance Computing. Available online: http://www.netapp.com/us/info/what-is-high-performance-computing.aspx. (accessed on 23 April 2020).
5. Henke, N.; Bughin, J.; Chui, M.; Manyika, J.; Saleh, T.; Wiseman, B.; Sethupathy, G. The age of analytics: Competing in a data-driven world. *McKinsey Glob. Inst.* **2016**, 1–28.
6. Singh, M.K. *Effective Big Data Management and Opportunities for Implementation*; IGI Global: Hershey, PA, USA, 2016.
7. Research and Markets. High Performance Computing (HPC) Market by Component, Infrastructure, Services, Price Band, HPC Applications, Deployment Types, Industry Verticals, and Regions 2020–2025. Available online: https://www.researchandmarkets.com/reports/4896466/high-performance-computing-hpc-market-by (accessed on 23 April 2020).
8. NVIDIA. Telecommunications Solutions for 5G Networks. Available online: https://www.nvidia.com/en-us/industries/telecommunications/. (accessed on 23 April 2020).
9. Ezell, S.J.; Atkinson, R.D. *The Vital Importance of High-Performance Computing to US Competitiveness*; Information Technology and Innovation Foundation: Washington, WA, USA,28 April 2016.
10. PRACE. Supercomputers for All; The Next Frontier for High Performance Computing SPECIAL REPORT. 2013. Available online: https://prace-ri.eu/wp-content/uploads/SupercomputersForAll.pdf (accessed on 23 April 2020).
11. Joseph, E.; Dekate, C.; Conway, S. *Real-World Examples of Supercomputers Used For Economic and Societal Benefits: A Prelude to What the Exascale Era Can Provide (Special Study)*; 2014. Available online: https://www.hpcuserforum.com/downloads/HPCSuccessStories.pdf (accessed on 15 May 2020)
12. Auweter, A.; Bode, A.; Brehm, M.; Huber, H.; Kranzlmüller, D. Principles of energy efficiency in high performance computing. In Proceedings of the International Conference on Information and Communication on Technology: Depok, Indonesia, Jun 24, 2020−Jun 26, 2020 pp. 18–25.
13. Halper, M. Supercomputing's super energy needs, and what to do about them. *Commun. ACM* **2015**, *9*, 93–99.
14. Flórez, E.; Pecero, J.E.; Emeras, J.; Barrios, C.J. Energy model for low-power cluster. In Proceedings of the 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), May 14, 2017, Madrid, Spain, pp. 1009–1016.
15. Amruta, M.K.; Satish, M.T. Solar powered water quality monitoring system using wireless sensor network. In Proceedings of the 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), March 22, 2013, Kerala, India, pp. 281–285.

16. Enterprise, H.P. Improving the energy efficiency of modern supercomputers. *Hewlett Packard Enterp. Dev. LP,* Springer: Cham, Switherland, **2017**, 1–9.
17. González, A. Trends in Processor Architecture. In *Harnessing Performance Variability in Embedded and High-performance Many/Multi-core Platforms*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 23–42.
18. Czarnul, P.; Proficz, J.; Krzywaniak, A. Energy-Aware High-Performance Computing: Survey of State-of-the-Art Tools, Techniques, and Environments. *Sci. Program.* **2019**, 19 (doi:10.1155/2019/8348791).
19. Matsuoka, S.; Endo, T.; Nukada, A.; Miura, S.; Nomura, A.; Sato, H.; Jitsumoto, H.; Drozd, A. Overview of TSUBAME3. 0 Green Cloud Supercomputer for Convergence of HPC AI and Big-Data. *E-Sci. J.* **2017**, *16*, 2–9.
20. Alsharif, M. H.; Nordin, R.; Ismail, M. Survey of green radio communications networks: Techniques and recent advances. *Journal of computer networks and communications* **2013**, *2013*, 13.
21. Letcher, C.W. *Green Computing-Desktop Computer Power Management at the City of Tulsa*; Oklahoma State University, Stillwater, Oklahoma, 2013.
22. Yi, G.; Loia, V. High-performance computing systems and applications for AI. *J. Supercomput.* **2019**, *75*, 4248–4251.
23. Lu, C.-P. AI, native supercomputing and the revival of Moore's Law. *APSIPA Trans. Signal Inf. Process.* **2017**, *6*.
24. Rong, H.; Zhang, H.; Xiao, S.; Li, C.; Hu, C. Optimizing energy consumption for data centers. *Renew. Sustain. Energy Rev.* **2016**, *58*, 674–691.
25. Council, N.R. *Getting up to Speed: The Future of Supercomputing*; National Academies Press: Cambridge, MA, USA, 2005.
26. Liao, X.-K.; Lu, K.; Yang, C.-Q.; Li, J.-W.; Yuan, Y.; Lai, M.-C.; Huang, L.-B.; Lu, P.-J.; Fang, J.-B.; Ren, J. Moving from exascale to zettascale computing: Challenges and techniques. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 1236–1244.
27. Strande, S.M.; Cai, H.; Cooper, T.; Flammer, K.; Irving, C.; von Laszewski, G.; Majumdar, A.; Mishin, D.; Papadopoulos, P.; Pfeiffer, W. Comet: Tales from the long tail: Two years in and 10,000 users later. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact, July 9–13, 2017, New Orleans, LA, USA, pp. 1–7.
28. Dayarathna, M.; Wen, Y.; Fan, R. Data center energy consumption modeling: A survey. *IEEE Commun. Surv. Tutor.* **2015**, *18*, 732–794.
29. Kant, K. Data center evolution: A tutorial on state of the art, issues, and challenges. *Comput. Netw.* **2009**, *53*, 2939–2965.
30. Bergamaschi, R.A.; Piga, L.; Rigo, S.; Azevedo, R.; Araújo, G. Data center power and performance optimization through global selection of p-states and utilization rates. *Sustain. Comput. Inform. Syst.* **2012**, *2*, 198–208.
31. Rizvandi, N.B.; Zomaya, A.Y. A Primarily Survey on Energy Efficiency in Cloud and Distributed Computing Systems. *arXiv* **2012**, arXiv:1210.4690.
32. Maiterth, M.; Koenig, G.; Pedretti, K.; Jana, S.; Bates, N.; Borghesi, A.; Montoya, D.; Bartolini, A.; Puzovic, M. Energy and power aware job scheduling and resource management: Global survey—Initial analysis, 2018, pp. 685–693. Available online: https://ieeexplore.ieee.org/abstract/document/8425478 (accessed on 24 April 2020).
33. Zamani, R.; Afsahi, A.; Qian, Y.; Hamacher, C. A feasibility analysis of power-awareness and energy minimization in modern interconnects for high-performance computing, 2007. Available online: https://ieeexplore.ieee.org/abstract/document/4629224/ (accessed on 24 April 2020).
34. Centre for Development of Advanced Computing (C-DAC) One-Day Symposium on Energy Efficiency Challenges for HPC Systems. Available online: https://www.cdac.in/index.aspx?id=pdf_Energy_Efficiency_Challenges_HPC_2019 (accessed on 24 April 2020).
35. Gupta, G. The Benefits of Bringing Artificial Intelligence to High Performance Computing. 2019. Available online: https://high-performance-computing.cioreview.com/cxoinsight/the-benefits-of-bringing-artificial-intelligence-to-high-performance-computing-nid-26530-cid-84.html (accessed on 24 April 2020).

36. Mei, X.; Wang, Q.; Chu, X. A survey and measurement study of GPU DVFS on energy conservation. *Digit. Commun. Netw.* **2017**, *3*, 89–100.
37. Johnsson, L.; Ahlin, D.; Wang, J. The SNIC/KTH PRACE prototype: Achieving high energy efficiency with commodity technology without acceleration, 2010, pp. 87–95. Available online: https://ieeexplore.ieee.org/abstract/document/5598259/ (accessed on 24 April 2020).
38. Kan, E.Y.; Chan, W.K.; Tse, T. EClass: An execution classification approach to improving the energy-efficiency of software via machine learning. *J. Syst. Softw.* **2012**, *85*, 960–973.
39. Saravanan, K.P.; Carpenter, P.M.; Ramirez, A. Power/performance evaluation of energy efficient ethernet (eee) for high performance computing., 2013, pp. 205–214, Available online: https://ieeexplore.ieee.org/abstract/document/6557171/ (accessed on 24 April 2020).
40. Dally, B. Power, programmability, and granularity: The challenges of exascale computing, Available online: https://ieeexplore.ieee.org/abstract/document/6139189/ (accessed on 24 April 2020).
41. Reed, D.A.; Dongarra, J. Exascale computing and big data. *Commun. ACM* **2015**, *58*, 56–68.
42. Hussain, S.M.; Wahid, A.; Shah, M.A.; Akhunzada, A.; Arshad, S. *Seven Pillars to Achieve Energy Efficiency in High Performance Computing and Big Data: An Application Perspective of Fog Computing*; Springer: Cham, Netherlands; 2019, 93–105.
43. Diouri, M.E.; Chetsa, G.L.T.; Glück, O.; Lefevre, L.; Pierson, J.-M.; Stolf, P.; Da Costa, G. Energy efficiency in high-performance computing with and without knowledge of applications and services. *Int. J. high Perform. Comput. Appl.* **2013**, *27*, 232–243.
44. Tan, L.; Song, S.L.; Wu, P.; Chen, Z.; Ge, R.; Kerbyson, D.J. Investigating the interplay between energy efficiency and resilience in high performance computing. In Proceedings of the 2015 IEEE International Parallel and Distributed Processing Symposium, May 25–29, 2015, Hyderabad, India, pp. 786–796.
45. Wang, Z.; Srinivasan, R.S. A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models, 2015, pp. 3438–3448. Available online: https://ieeexplore.ieee.org/abstract/document/7408504 (accessed on 24 April 2020).
46. Dounis, A.I. Artificial intelligence for energy conservation in buildings. *Adv. in Build. Energy Res.* **2010**, *4*, 267–299.
47. Jiang, Z.; Gao, W.; Wang, L.; Xiong, X.; Zhang, Y.; Wen, X.; Luo, C.; Ye, H.; Lu, X.; Zhang, Y. HPC AI500: A benchmark suite for HPC AI systems, 2018, pp. 10–22, Available onlind: https://link.springer.com/chapter/10.1007/978-3-030-32813-9_2 (accessed on 24 April 2020).
48. García Martín, E. Energy efficiency in machine learning: A position paper. In Proceedings of the 30th Annual Workshop of the Swedish Artificial Intelligence Society SAIS, May 15 and 16, 2017, Karlskrona, Sweden, pp. 68–72.
49. IntelCorporation. Bringing AI Into Your Agency HPC Environment. Available online: https://www.govexec.com/media/intel_ai-hpc_eguide.pdf (accessed on 30 April 2020).
50. Imes, C.; Hofmeyr, S.; Hofmann, H. *Energy Efficiency in HPC with Machine Learning and Control Theory*, 2017, Available online: https://sc17.supercomputing.org/SC17%20Archive/tech_poster/poster_files/post215s2-file3.pdf (accessed on 30 April 2020).
51. Beloglazov, A.; Buyya, R.; Lee, Y.C.; Zomaya, A. A taxonomy and survey of energy-efficient data centers and cloud computing systems. In *Advances in Computers*; Elsevier: Amsterdam, The Netherlands, 2011; Volume, 82, pp. 47–111.
52. Vetter, J.S.; Mittal, S. Opportunities for nonvolatile memory systems in extreme-scale high-performance computing. *Comput. Sci. Eng.* **2015**, *17*, 73–82.
53. Freeh, V.W.; Lowenthal, D.K.; Pan, F.; Kappiah, N.; Springer, R.; Rountree, B.L.; Femal, M.E. Analyzing the energy-time trade-off in high-performance computing applications. *IEEE Trans. Parallel Distrib. Syst.* **2007**, *18*, 835–848.
54. Labasan, S. *Energy-Efficient and Power-Constrained Techniques for Exascale Computing*; Semanticscholar: Seattle, WA, USA, 2016.
55. Wlotzka, M.; Heuveline, V. Energy-efficient multigrid smoothers and grid transfer operators on multi-core and GPU clusters. *J. Parallel Distrib. Comput.* **2017**, *100*, 181–192.
56. Graham, S.L.; Snir, M.; Patterson, C.A. Bolstering US Supercomputing. *Issues Sci. Technol.* **2005**, *21*, 28–32.
57. Kelechi, A.H.; Alsharif, M.H.; Ramly, A. A.; Abdullah, N.F.; Nordin, R. The Four-C Framework for High Capacity Ultra-Low Latency in 5G Networks: A Review. *Energies* **2019**, *12*, 3449.

58. Alsharif, M.H.; Nordin, R. Evolution towards fifth generation (5G) wireless networks: Current trends and challenges in the deployment of millimetre wave, massive MIMO, and small cells. *Telecommunication Systems* **2017**, *64*, 617-637..
59. Baldemair, R.; Dahlman, E.; Fodor, G.; Mildh, G.; Parkvall, S.; Selen, Y.; Tullberg, H.; Balachandran, K. Evolving wireless communications: Addressing the challenges and expectations of the future. *IEEE Veh. Technol. Mag.* **2013**, *8*, 24–30.
60. ITU-R. 5G —Fifth Generation of Mobile Technologies. Available online: https://www.itu.int/en/mediacentre/backgrounders/Pages/5G-fifth-generation-of-mobile-technologies.aspx. (accessed on 30 April 2020).
61. Durisi, G.; Koch, T.; Popovski, P. Toward massive, ultrareliable, and low-latency wireless communication with short packets. *Proc. IEEE* **2016**, *104*, 1711–1726.
62. A. Morris. Scaling for 5G: From Data Centers to the Edge. Available online: https://www.hpcwire.com/solution_content/ibm/scaling-for-5g-from-data-centers-to-the-edge/ (accessed on 24 April 2020).
63. Dahlman, E.; Mildh, G.; Parkvall, S.; Peisa, J.; Sachs, J.; Selén, Y. 5G radio access. *Ericsson Rev.* **2014**, *6*, 1−28.
64. Fu, Y.; Wang, S.; Wang, C.-X.; Hong, X.; McLaughlin, S. Artificial intelligence to manage network traffic of 5G wireless networks. *IEEE Netw.* **2018**, *32*, 58–64.
65. Levis, B. Scaling HPC for 5G, AI, and Whatever's Next. Available online: https://www.insight.tech/content/scaling-hpc-for-5g-ai-and-whatever-s-next (accessed on 25 April 2020).
66. Al-Quzweeni, A.N.; Lawey, A.Q.; Elgorashi, T.E.; Elmirghani, J.M. Optimized energy aware 5G network function virtualization. *IEEE Access* **2019**, *7*, 44939–44958.
67. Alsharif, M.H.; Kelechi, A.H.; Albreem, M.A.; Chaudhry, A.C.; Zia, M.S.; Kim, S. Sixth Generation (6G) Wireless Networks: Vision, Research Activities, Challenges and Potential Solutions. *Symmetry* **2020**, *12*, 676.
68. Zhou, Y.; Li, N.; Li, H.; Zhang, Y. Regression cloud models and their applications in energy consumption of data center. *J. Electr. Comput. Eng.* **2015**, *2015*, 9.
69. Shoukourian, H.; Wilde, T.; Auweter, A.; Bode, A. Monitoring power data: A first step towards a unified energy efficiency evaluation toolset for HPC data centers. *Environ. Model. Softw.* **2014**, *56*, 13–26.
70. Chen, Y.-L.; Chang, M.-F.; Yu, C.-W.; Chen, X.-Z.; Liang, W.-Y. Learning-Directed Dynamic Voltage and Frequency Scaling Scheme with Adjustable Performance for Single-Core and Multi-Core Embedded and Mobile Systems. *Sensors* **2018**, *18*, 3068.
71. Trestian, R. *Next-Generation Wireless Networks Meet Advanced Machine Learning Applications*; IGI Global, Hershey, PA, USA, 2019.
72. Alsharif, M.H.; Kelechi, A.H.; Yahya, K.; Chaudhry, S.A. Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends. *Symmetry* **2020**, *12*, 88.
73. Osisanwo, F.; Akinsola, J.; Awodele, O.; Hinmikaiye, J.; Olakanmi, O.; Akinjobi, J. Supervised machine learning algorithms: Classification and comparison. *Int. J. Comput. Trends Technol. (IJCTT)* **2017**, *48*, 128–138.
74. Wang, Z.; Srinivasan, R.S. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renew. Sustain. Energy Rev.* **2017**, *75*, 796–808.
75. Wang, T.; Xia, Y.; Muppala, J.; Hamdi, M. Achieving energy efficiency in data centers using an artificial intelligence abstraction model. *IEEE Trans. Cloud Comput.* **2015**, *6*, 612–624.
76. Chen, K.; Lin, G. Optimization of multiple-module thermoelectric coolers using artificial-intelligence techniques. *Int. J. Energy Res.* **2002**, *26*, 1269–1283.
77. Lee, J.; Stanley, M.; Spanias, A.; Tepedelenlioglu, C. Integrating machine learning in embedded sensor systems for Internet-of-Things applications. 2016, pp. 290–294, Available onlind: https://ieeexplore.ieee.org/abstract/document/7886051/ (accessed on 30 April 2020).
78. Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms, 2016, pp. 1310–1315, Available online: https://ieeexplore.ieee.org/abstract/document/7724478/ ((accessed on 30 April 2020).
79. Chen, S.H.; Jakeman, A.J.; Norton, J.P. Artificial intelligence techniques: An introduction to their use for modelling environmental systems. *Math. Comput. Simul.* **2008**, *78*, 379–400.
80. Belhaj, S.; Tagina, M. Modeling and prediction of the internet end-to-end delay using recurrent neural networks. *J. Netw.* **2009**, *4*, 528–535.

81. Buskirk, T.D.; Kirchner, A.; Eck, A.; Signorino, C.S. An introduction to machine learning methods for survey researchers. *Surv. Pract.* **2018**, *11*, 2718.
82. Rodvold, D.; McLeod, D.; Brandt, J.; Snow, P.; Murphy, G. Introduction to artificial neural networks for physicians: Taking the lid off the black box. *Prostate* **2001**, *46*, 39–44.
83. Jain, A.K.; Mao, J.; Mohiuddin, K.M. Artificial neural networks: A tutorial. *Computer* **1996**, *29*, 31–44.
84. Mike Rycroft. Energy Management in Industry: Can AI Improve Energy Efficiency? Available online: https://www.ee.co.za/article/energy-management-in-industry-can-ai-improve-energy-efficiency.html (accessed on 25 April 2020).
85. Sui, X.; Yang, Y.; Xu, X.; Zhang, S.; Zhang, L. The sampled-data consensus of multi-agent systems with probabilistic time-varying delays and packet losses. *Phys. A Stat. Mech. Appl.* **2018**, *492*, 1625–1641.
86. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.-C.; Kim, D.I. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3133–3174.
87. Li, J.; Gao, H.; Lv, T.; Lu, Y. Deep reinforcement learning based computation offloading and resource allocation for MEC. 2018, pp. 1−6, Available online: https://ieeexplore.ieee.org/abstract/document/8377343/ (accessed on 25 April 2020).
88. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 67.
89. Binsahaq, A.; Sheltami, T.R.; Salah, K. A survey on autonomic provisioning and management of QoS in SDN networks. *IEEE Access* **2019**, *7*, 73384–73435.
90. Gao, J. *Machine Learning Applications for Data Center Optimization*; Semanticscholar: Seattle, WA, USA, 2014.
91. Evans, R.; Gao, J. Deepmind AI reduces Google data centre cooling bill by 40%. *DeepMind Blog* **2016**, *20*, 158.
92. DeepMind, A. *Reduces Google Data Centre Cooling Bill by 40%*; Semanticscholar: Seattle, WA, USA, 2016.