Thermal Science and Applications
https://www.sciltp.com/journals/tsa

Review

Utilizing Artificial Intelligence Techniques for Cooling Optimization and Energy Management in Smart Data Centers

Mychal Berlinski and Behnaz Rezaie *

Physical and Computational Science Division, University of Pittsburgh, 300 Campus Drive, Bradford, PA 16701, USA
* Correspondence: brezaie@pitt.edu

How To Cite: Berlinski, M.; Rezaie, B. Utilizing Artificial Intelligence Techniques for Cooling Optimization and Energy Management in Smart Data Centers. *Thermal Science and Applications* **2026**, *1*(3), 253–279. https://doi.org/10.53941/tsa.2026.100016

Received: 6 August 2025
Revised: 15 July 2026
Accepted: 23 July 2026
Published: 14 August 2026

Abstract: The global shift to digital technologies has spurred a surge in data centers, intensifying server room cooling challenges. Effective cooling ensures optimal IT equipment operation, reliability, longevity, and data security, requiring stable, efficient environments for critical operations. Traditional HVAC systems fall short for data centers, prompting specialized advancements, hybrid cooling integrations, and optimization techniques. Artificial Intelligence (AI), like Artificial Neural Networks (ANN), revolutionizes server room cooling by training hardware and systems for minimal energy use. This study reviews recent advances in smart data center cooling via AI-driven HVAC, dynamic control, environmental monitoring, adaptive vents, thermal mapping, and data-driven strategies. AI integrates with building management systems for facility-wide coordination, enhancing efficiency. Key AI capabilities include real-time monitoring, predictive analysis, and adaptive responses to dynamic conditions, yielding higher cooling efficiency, reduced energy consumption, extended hardware life, and improved reliability. These enable sustainable, cost-effective operations essential for our digital era. Results highlight AI's role in temperature and load prediction, enabling proactive management. Various ANN techniques, detailed here, balance loads and temperatures, optimizing server room performance through intelligent control. The future research roadmap for cooling data centers concludes in the development of smart HVAC systems, data analytics, and AI-based management strategies, energy consumption optimization, interdisciplinary research, and adaptive environment and load regulation.

Keywords: AI-driven data center cooling; artificial neural networks HVAC; smart data center thermal management; predictive cooling optimization; energy-efficient server room cooling; machine learning building management systems

1. Introduction

Thermal management of the data centers in the digital world, which we are moving toward, is crucial to support the reliability of server rooms and the necessary IT infrastructure. Advancing the conventional HVAC system for cooling the server room has been the original approach.

Conventional HVAC-based approaches for data center and building cooling have long focused on improving energy efficiency through system-level optimization and waste heat recovery. For example, in response to future EU carbon emission regulations, Montagud-Montalvá et al. investigated experimental strategies for recycling waste heat through advanced district heating and cooling systems in Valencia, Spain [1]. Using CYPETHERM HE Plus, the authors modeled selected campus buildings based on internal loads, thermal conditions, geometry, and construction materials. Waste heat from a sports facility and a data center was recovered to support heating and cooling demands, with surplus energy stored in on-site thermal energy storage. Three operational scenarios were evaluated, yielding carbon savings ranging from 64,035 kg to 534,780 kg of CO_2. While the most feasible scenario achieved a 64% contribution to building energy demand with moderate emissions reduction, broader implementation was constrained by infrastructure reconfiguration requirements and operational complexity, highlighting the limitations of static, design-intensive solutions.

Similarly, Zhang et al. examined hybrid thermosyphon–refrigeration systems to improve data center cooling performance across diverse climatic conditions in China [2]. Although thermosyphons and free cooling systems offer energy savings during cooler seasons, their effectiveness diminishes in warmer climates. By integrating mechanical refrigeration, the proposed system enabled year-round operation and reduced energy consumption, with savings ranging from 5.4% in Guangzhou to 47.3% in Harbin. Despite these improvements, system performance remained highly climate-dependent, underscoring the limited adaptability of conventional and hybrid cooling technologies.

Collectively, these studies demonstrate meaningful progress in improving cooling efficiency through hardware and system-level innovations, yet they also reveal inherent limitations related to climate sensitivity, infrastructure constraints, and static control strategies. These challenges have motivated the growing adoption of AI-driven approaches capable of dynamic monitoring, prediction, and adaptive control. Building on this evolution, the present review focuses on AI-based cooling strategies for data centers, examining how intelligent methods address the shortcomings of conventional HVAC systems and outlining emerging opportunities for energy-efficient, scalable, and resilient thermal management.

Existing data center cooling research has primarily emphasized hardware-based and system-level solutions such as waste heat recovery, thermosyphon-assisted cooling, and hybrid refrigeration systems. While these approaches can reduce energy consumption and emissions, they are often limited by climate dependency, high capital costs, and a lack of adaptability to dynamic server loads. This literature review identifies AI, like ANNs, as a promising pathway to overcome these limitations by enabling predictive, adaptive, and real-time cooling control. However, challenges remain, including data availability, model generalizability, integration with legacy HVAC systems, and practical implementation for operators. By synthesizing existing studies, this review presents a clear roadmap for AI-driven data center cooling, highlighting both the barriers that must be addressed and the opportunities for scalable, energy-efficient, and reliable thermal management in next-generation data centers. These benefits contribute to more sustainable and cost-effective data center operations, which are essential in our increasingly digital and connected world. Providing insight for the field engineers and researchers. Figure 1 shows the approaches for cooling the server rooms that are investigated in this study.

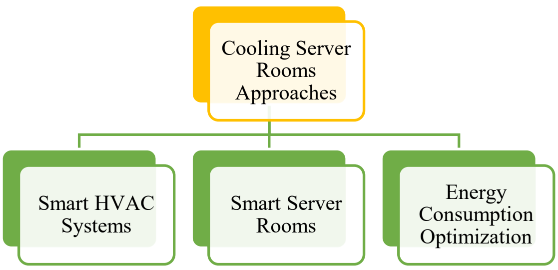

Figure 1. Major strategies for data centers.

2. Smart HVAC Systems

The reliability of server rooms is greatly impacted by cooling them. Therefore, customizing the conventional HVAC systems for adjusting the server rooms' conditions has been a common approach for cooling data centers. AI can integrate with the overall building management system and HVAC infrastructure to coordinate cooling efforts across the entire facility by focusing on control systems, ensuring efficient and coordinated cooling operations.

Opalic et al. studied the ability of CO_2 as a refrigerant to cool a smart warehouse to cut down on costs for maintenance and operation [3]. A model was developed based on an ANN, hidden layer (HL), to monitor and predict the ambient warehouse temperature and humidity. The model was designed to measure and calculate the amount of electrical consumption and CO_2 mass flow rate. The PV panels were employed to power the ANN and other cooling components, such as the electrical boiler, batteries, chilled storage spaces, and the overall cooling system. Any excess energy and waste heat could be transferred to TES. Inputs to the ANN and other components of the system included cooling capacity, electrical power, electrical current, and mass flow rate. The experiment was designed to measure the power and compare the mean average percent errors and mean squared error. It was observed that the mean average percent error (MAPE) of the accuracy of the ANN in the case study ranged from 5–12% during the case-study segment. In a comparison of measured and modeled power usage, the cooling system operated between 74.9 and 98.3 bar, and it was found that the MAPE was calculated at 3.13% and 1.87% for the ANN outputs for power meters and frequency-controlled compressors, respectively. It was also noted that an increase in data quality could increase accuracy for a MAPE of 1.76–1.87% for flow and power, respectively.

Zhao and Liu explored the impact of artificial intelligence on cooling systems for office buildings by employing a new method for AI, called the Wavelet-PLS-SVM and it was compared to PLS and SVM, and a combination of Wavelet-PLS-SVM [4]. Input parameters for this study included external disturbances, referring to environmental by related changes in weather, internal disturbances referring to occupancy, and the lighting in the selected area of the building. Data was gathered from a building in Tianjin, China. It was found that Wavelet-PLS-SVM reduced the error in prediction accuracy of heating and cooling loads by 5.70% compared to Wavelet-SVM and 5.53% compared to PLS. Furthermore, the Wavelet-PLS-SVM enhanced heating predictions by 5.46% compared to the SVM method.

Lee et al. studied the ability of artificial intelligence to predict control measures for TES systems for different cooling and heating loads by utilizing model predictive control (MPC) as part of the optimization of cooling operations [5]. Experimentation compared MPC and rule-based control (RBC) in which the model had a more positive effect when paired with AI. During testing, the team found that MPC removed heat 7.4% more than the RBC, in a cumulative study. In a high load condition, the MPC model reduced overall operation costs by 9.06%. On medium-load cooling operations, the MPC saved 9.29% of operation costs. With a TES, during a high cooling load test, the MPC could further reduce costs by 14.56%.

The competence of artificial intelligence to efficiently manage energy consumption for cooling was studied by the use of a deep learning technique and swapping a physical sensor for a Recurrent Neural Network (RNN) [6]. RNN was compared to an ANN, adaptive neuro-fuzzy inference system (ANFIS), and a 4-layered ANN. A set of experiments was conducted with controllable loads and the RNN outperformed all other systems with input parameters that considered return air volume, air supply volume, pressure set-point, temperature set-point, and air supply temperature. For heating and cooling split systems the RNN balanced 84.21% of the load, while the ANN, ANFIS, and 4-layered ANN balanced 79.62%, 81.13%, and 82.66%, respectively. For the efficiency of balancing the loads, the RNN, also ranked high for the heating and cooling split systems, measuring 73.934%, while the ANN, ANFIS, and 4-layered ANN measured 69.979%, 71.284%, and 72.601%, respectively.

Hosoz et al. studied artificial neural networks to predict the performance of cooling towers [7]. The team used 70% of the collected data to train the ANN model and 30% to test the model for predictiveness, with input data being dry bulb temperature, the relative humidity for air streams in the cooling towers, incoming water stream temperatures, and water and air mass flow rates. The anticipated outputs of the experiment were prediction accuracies of heat rejection rates, water evaporation rate, water temperature leaving the tower, air dry bulb temperature, and relative humidities of the air streams. The results indicated that for the heat rejection rate vs. experimental value, the accuracy was 99.81%. Other measured accuracy rates included for water evaporated were 99.78%, relative humidity at 99.99%, water temperature leaving the tower at 99.99%, and air dry bulb temperatures at 99.99%.

Yigit and Ertunc developed and studied the use of neural networks and their ability to predict the temperature and humidity of exiting cooling coils as part of a cooling system, to help prevent and minimize the formation of frost at the outlet and increase the efficiency of the cooling system [8]. The inputs to the ANN were the coolant temperature, ranging from −25 °C to 0 °C, air inlet temperature ranging from 25 °C to 45 °C, air humidity, ranging from 20% to 50%, and air velocity, ranging from 1.5 m/s to 11 m/s, in which, with anticipated output parameters being measured temperature and humidity levels of the air stream leaving the cooling coils. 70% of the selected data was meant to train the ANN and 30% to test the ANN. Results were found that the correlation between predicted temperature and experimental temperatures was 99.9% accurate, with the humidity, the accuracy was measured and calculated to be 98.2%.

Li et al. proposed an ANN prediction model for cooling loads of a building [9]. The study was designed based on the improvement of the interpretability of RNN to predict cooling loads 24 h ahead. The setup selected used the multi-input multi-output (MIMO) method. The inputs were past and historical time series datasets while designated output parameters included predictions for building energy consumption accuracy. Two variations of RNN algorithms were observed, long short-term memory (LSTM) and gated recurrent units (GRU). Different trials resulted in comparisons of LSTM and GRU, they resembled each other in terms relating to the coefficient of the variation of the room mean square error (CVRMSE) percentages. For the trials with mechanism attention, the LSTM and GRU had 32.8% and 31.5% CVRMSE, respectively. Without the mechanism attention, the LSTM and GRU measured in at 35.2% and 39% with CVRMSE, respectively. During experimentation, an epoch was defined as a complete, full pass through the training datasets. Training times per epoch also decreased, with initial LSTM and GRU values being 8.8 s and 7.6 s, being improved to 6.6 s for LSTM and 5.3 s for GRU.

Hou et al. examined artificial neural networks model rough set theory (RS) methods to predict cooling loads in buildings [10]. The model was developed with the combination of theory RS and ANN to create the multi-RSAN model (MRAN). The inputs to the model were historical datasets, with a focus on airflow rates, relative humidity, air pressure, and air temperature. Focused output parameters included prediction accuracies of cooling loads and their respective energy consumption estimations. The experiment consisted of the simulation of the MRAN model being compared to other models, such as the autoregressive integrated moving average (ARIMA) model. The maximum and mean error for MRAN were 9.50% (45.7 kW) and 3.65% (18.8 kW). These values were the overall and general lowest in the group, compared to ARIMA and RSAN 1–6.

Kamel et al. studied energy consumption prediction models for residential buildings [11]. The proposed model was developed based on data-driven models (DDM) in which synthetic data was used. Different methods were as follows for developing the prediction models: M0 used the Filter Method (Pearson correlation) and linear regression, M1 used the Wrapper Method (backward elimination) and ordinary least square, M2 used the wrapper method (recursive feature elimination) and linear regression, M3 used embedded method (Lasso Regularization), M4 used XGBoost method, and M5 used linear regression on the previous M4 features. For inputs generated by the Building Energy Optimization Tool (BEopt) for the heating, cooling, ventilation, and hot water temperatures. The output parameters were total energy used for individual heating or cooling load data-driven predictions. Through simulation, it was noted that cooling energy consumption was irrelevant from October to May while the ventilation load was irrelevant for October through December and April and May, due to seasonal changes and intermittent periods of use during these periods for the specific functions. Trials consisted of six different methods of DDMs, each scoring incredibly high. In order as listed, the accuracy was 99.7%, 99.8%, 99.8%, 99.8%, 99.9%, and 99.7% for prediction of cooling loads. It was concluded that the DDMs, in total, have a 99.7% accuracy in prediction rating overall for heating and cooling loads.

The prediction model for cooling and heating loads in office buildings was investigated and developed [12]. The study was designed by focusing on the demand-driven cooling control system integrated with learning techniques into the existing system. Inputs for the model included days of the week, arrival times and building occupation, and idle server temperatures while the output was the prediction of energy consumption and its accuracy. Experimentation specifically focused on the machine learning algorithm, k-nearest neighbor (KNN) as the prediction algorithm. It was discovered that during two evaluations, machine learning predicted with an accuracy of 90% in evaluation 1, and 95% in evaluation 2, both regarding energy consumption. Only 7% of energy was saved for multi-person offices, due to the dense occupancy and higher demand for cooling energy. For medium-sized office spaces, 21% of energy was reduced, and for meeting rooms, the saved energy was in the range of 25% to 52%.

Sha et al. explored a prediction model to reduce energy consumption and improve cooling efficiency by utilizing machine learning techniques on mechanical ventilation for large buildings [13]. The study was designed to implement and compare nine different machine learning prediction techniques including linear regression (LR), ridge regression (RR), elastic net (ELN), support vector machine regression (SVR), gradient tree boosting (GTB), multilayer perceptron (MLP), RNN, LSTM, a persistence model based on a naïve prediction method, and convolutional neural network (CNN). Python, Scikit-Learn, MATLAB, and Keras programs were used for developing the models. Inputs for the models included cooling load, water temperature, correlation coefficients, and power consumed to generate a prediction for energy consumption. Through different trials, it was discovered that GTB produced the lowest CVRMSE at 12.3% while the persistence model had the highest at 39.4%. Overall, GTB retained the lowest error rates and had the highest accuracy of the models selected to predict energy consumption, with CVRMSE values of 12.3%, 12.4%, and 12.7% over the periods of 1 h, 30 min, and 6-min run time, respectively. Also noted were the normalized mean bias error (NMBE) values of the GTB measured to be 1.2%, 0.5%, and −0.2% over the periods of 1 h, 30 min, and 6 min, respectively. It was also noted that the results were affected by the weather conditions. The optimal air flow rate for ventilation was 213.9 m^3/s while for the worst future scenario related to weather conditions, the optimal flow rate was 107 m^3/s. Factors that influence the flow rate due to weather are dry-bulb temperatures and humidity levels.

Ma et al. investigated a predictive model for TES in cooling systems for buildings [14]. The model consisted of four main components: the TES tank, chillers cooling towers, and the physical building model. The model was structured around model-based predictive control (MPC). Inputs to the model included the date, time of day, local weather, electrical energy price, and occupancy of the selected building. The model aimed to generate a prediction of the total energy required for the cooling system. Results of the simulation demonstrated that the MPC reduced electrical costs by \$1265 and improved coefficient of performance (COP) by 1.5% in the feedback loop. Another iteration saved \$1205 and increased the COP by 11.9%.

Zabala et al. developed a predictive model using Python and Modelica algorithms in predicting and developing efficient district cooling systems with the goal for the study to develop a predictive model capable of complying with MPC [15]. The MPC captured significant dynamic changes in physical model variables, with local weather changes having an effect, with the model also being economically intelligent to reduce costs. The model was enhanced by applying synthetic data, autoregressive exogenous (ARX) input, and state space (SS) model representation. Inputs included with the enhancement included the outlet temperature of chilled water, chilled water flow rate, inlet temperature of the chilled water, and inlet temperature of cooling air, with outputs measuring outlet temperatures of chilled water and consumed electrical power. During analysis, it was discovered that mean supply temperatures could increase by 81.3%. The minimum increase in improvements was −3.33%, while maximum improvements measured as high as 98.59%. The MPC was also able to generate a mean energy savings of 50.09%, with the minimum and maximum energy savings being 6.70% and 54.10% respectively. It was noted that the MPC's effectiveness and performance depend on the model's reliability and data selected to be tested.

Xu and Wang explored artificial neural networks to manage solar cooling systems by focusing on the relationship between ANNs and solar compound parabolic collectors (CPC) [16]. Solar CPCs utilize the thermal energy provided by the sun to create heat transfer, generating efficiencies of 50–48% when operating between 130 °C and 150 °C TRaNsient System Simulation program (TRNSYS) was used to simulate the solar CPC cooling system while for the simplicity of the algorithm Back Propagation (BP) neural network was utilized. Inputs to the simulation included heat source temperature, cooling water inlet temperature, required cooling, and chilled water inlet temperature. Output parameters included cooling power and coefficient of performance of the cooling system. Through various simulations, it was determined that when storage tank volume increased from 1.0 m^3 to 4.5 m^3, the efficiency increased from 43.4% to 47.9%, when cut-off temperatures increased for the chillers from 90 °C to 115 °C, the solar CPC efficiency decreased from 47% to 37.7%.

The cooling of an overclockable control process unit (CPU) was proposed as a combination of air, water (cold plates), and boilerplates by determining their effectiveness in cooling [17]. Overclocking was the customization of computer or server settings to increase performance by tweaking power levels, voltages, memory settings, and frequencies. During overclocking, it was essential to ensure necessary cooling was provided due to the change in settings could change heat exchange rates that were not affected during normal operation. To observe the cooling of CPUs, a test was set up comparing an air-cooled, cold plate, and 2-phase immersion boilerplate The team used an Intel i9 9900k CPU with an Intel XTS 100H commercially available air-cooled heat sink for the air-cooled section, Indium, a dielectric fluid, HFE 7000, and grease for the cold plates and 2 phase immersion. Parameters considered during the study of these cooling applications included the inlet temperature of the different coolants, the inlet flow rate of the different coolants, and thermal interface materials (TIM). The frequency of the CPU was tested in different conditions. The two-phase cooling system was studied with a surrounding air temperature of 34 °C. It was observed that 34 air-cooled methods could cool at stable frequencies of 3.3 GHz, 4 GHz, and 3.6 GHz when using Prime95, XTU, and Cinebench datasets and workloads, respectively. When workloads such as Prime95, XTU, and Cinebench were used, CPU power was found to increase by 38%, 42%, and 48% using cold plates. The highest efficiency cooling method was the 2-phase cooling using HFE 7000 with Indium, which produced frequencies of 110% higher than air-cooled and 50% higher than cold plates.

3. Smart Server Rooms

The goal of a smart server room is to create an environment that operates efficiently, proactively addresses issues, and reduces the need for manual intervention. These technologies lead to cost savings, increased uptime, improved security, and better resource utilization for businesses that rely on their IT infrastructure. These are achievable through two major parallel paths, IT infrastructure management and cooling management of the data centers. Figure 2 illustrates different approaches for heat management in smart server rooms.

3.1. Managing Data Centers' Schedules

Work scheduling in server rooms is of paramount importance for ensuring the smooth and efficient operation of IT infrastructure and maintaining the reliability of critical services. Work scheduling in server rooms is crucial for maintaining reliability, optimizing resource usage, reducing downtime, ensuring safety, and facilitating efficient management of IT infrastructure. It allows organizations to balance the need for updates and maintenance with the requirement for uninterrupted service delivery. Using ANN methods has been a good approach for researchers to achieve their goals.

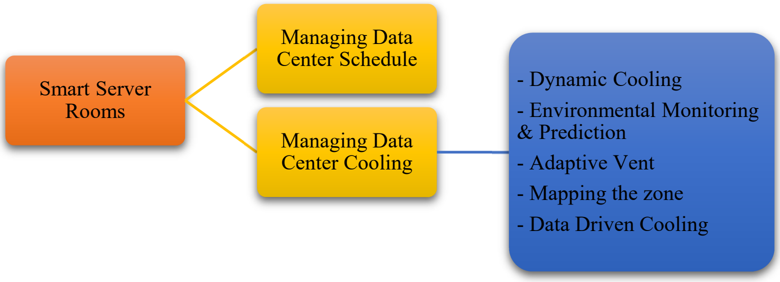

Figure 2. Heat management in data centers.

A team of researchers studied different machine-learning methods for optimizing memory systems and memory performance for server machines [18]. Methods such as the nearest neighbor (NN), naïve Bayes (NB), decision tree (DT), Ripper classifier (RP), SVM, LR, MLP, and radial basis function (RBF) were compared were used in the study. Each machine learning method was not tested against each other, but in the instances, some were, DT was found to perform at a higher geomean when considering the parameters of different benchmarking configurations. The weighted geomeans determined the best optimization for prefetching benchmarking categories such as ContentAds, Clustering, PageScorer, ContentIndexer, Sorting, and KeywordCluster. Also, DT had a weighted geomean of 0.989, while LR and MLP had weighted geomeans of 0.978 and 0.988, respectively. During speedup tests, the overall best-suited method was determined to be DT, due to the simple tree structure models with speedup times of 99% compared to the most optimal configuration that was compared.

Liang et al. researched the utilization of Advantage Actor-Critic (A2C) deep reinforcement learning to predict job scheduling for servers [18]. The model used A2cScheduler to manage the job scheduling, reducing estimation variances and improving parameter efficiency. In conjunction with the TensorFlow machine learning framework, the A2cScheduler could be implemented to return data in a Bernoulli process manner. The A2cScheduler algorithm was tested on slowdown, waiting times, and average completion times, and also compared against Tetris and Tetris and Shortest Job First (SJF) algorithms. Compared to SJF, the A2cScheduler on average had a lower slowdown after 500 episodes. Respectively, the slowdown job traces were 1.82, 1.61, and 1.01. It was noted that A2cScheduler has higher performance for the slowdown, average completion time, and average waiting time, ranging from 0.6 to 0.9. It was found that the A2cScheduler measured a waiting time of 0.01, while Tetris and SJF measured 1.25 and 1.21, respectively. Completion times measured 2.58 for A2cScheduler while Tetris and SJF measured 5.55 and 5.51, respectively.

Wei et al. explored the utilization of deep reinforcement learning (DRL) in the optimization of resource allocation for data centers by applying the natural evolution strategy-based DRL by means of multi-dimensional resource allocation (NESRL-MRM) to balance the utilization of resources across multiple dimensions to optimize the available resources [19]. Inputs to the NESRL-MRM included the dataset for tasking, the learning rate, and noise standard deviation. Python 3.8 as a visual interface, often called an integrated development environment (IDE), and implemented the NESRL-MRM by use of TensorFlow were used for trials. The NESRL-MRM was compared to DRL baseline schemes such as A2cScheduler, MinDIFF, tight packing heuristic algorithm, and round-robin resource allocation (RR). It was concluded that NESRL-MRM performs superiorly than other algorithms when simulated using Google and Amazon EC2 datasets. NESRL-MRM achieved 40% blocking heuristics, with 20% being in medium and high workloads using Google data. Compared to A2c, the NESRL-MRM had a time efficiency improvement of 85%, using a balanced use of resources.

Gebreyesus et al. investigated the optimization of data centers by use of a machine learning method and Shapley Additive exPlanation (SHAP) algorithm [19]. This algorithm was compared with other methods such as TreeSHAP with similar algorithms and faster computing than other methods. Python was used to process the information that was conveyed from the SHAP into virtual graphs and organizational tables. Machine learning techniques such as TreeSHAP, random forest (RF), and extreme gradient boosting (XGB) were used to generate models for trials. The experiment results confirmed that SHAP-assisted methods performed superior to other methods measuring performance in terms of mean absolute error (MAE), MAPE, and root mean squared error (RSME). The performance of the SHAP-assisted measured to have a 0.401 MAE, 0.245 RMSE, and 0.0035 MAPE.

Fernández-Cerero et al. studied the effects of employing a new resource management model by employing machine learning, specifically using gradient-boosting regression models with SCORE simulation tool [20]. The datasets from Google were obtained to train the regression model. Inputs to the model included inter-arrival times, task numbers, duration of a job, and resource usage. Through various trials, it was learned that using the regression model 80% of the jobs were scheduled in less than ~270 s each time a decision was made every 120 s while 80%

of the jobs were scheduled in less than 200 s when a decision was made for each job scheduling action. It was concluded that the regression model was trained with higher than 80% accuracy.

A research team investigated the use of matrix prediction on data centers using a deep belief network (DBN) architecture approach with a regression model [21]. Restricted Boltzmann Machine (RBM) was a crucial part of the DBN due to the Gaussian and Bernoulli distributions and the frameworks creating a series of hidden and visible layers. Experiments were conducted to compare the DBN approach and other modern algorithms such as SRMF, TomoGravity, and principal component analysis (PCA). The results of the experiment yield an improvement in error by 32.7% against SRMF, 23.4% against TomoGravity, and 23.4% against PCA in the first iteration for estimation. In the second iteration, the results were 47.7%, 45.7%, and 72.0% reduction in error for estimation, respectively.

Levi et al. developed a model coupled with Simulink to develop models for simulations, ultimately creating a neural network-esque prediction method to predict characteristics in a data center [22]. Inputs to the neural network included the total workload and amount of computation nodes. Then neural network (NN) with two other forms of modeling, nonlinear regression (NR) and SVM were compared by testing the three forms of modeling with run times, maximum waiting times, and energy consumption. For run times, NR measured 119.81%, SVM measured 37.99% and the NN measured 21.00% in mean average percent errors (MAPE). For waiting times MAPE, NR measured 203.06%, SVM measured 50.61%, and NN measured 30.43%. For energy consumption MAPE, NR measured 2.82%, SVM measured 50.61%, and NN measured 0.95%.

Guo et al. studied the use of immersion cooling for servers by employing a dielectric fluid to immerse the machine or chips to maintain cooling properties [23]. The hydrocarbon-based (mineral oils or synthetic oils) and fluorocarbons (engineered fluids) were studied which were used for cooling applications such as single-phase and double-phase cooling. For single-phase cooling, the fluid was operating between −9 to 74 °C with a power supply of 600 W. To select the liquid, Chevron's PAO-2 was implemented. The highest temperature a component reached during the experiment was 104 °C while the environmental temperature was 45 °C. It was determined that to increase cooling capabilities, the CPU would require to be optimized in a high-viscosity fluid.

The capability of novel passive cooling systems on temperature distribution in data centers was researched by Wang et al. using liquid nitrogen (LN_2) and finned tube heat exchangers (FTHEs) for cooling data centers [24]. Throughout the experiment, FTHEs were observed and information regarding inlet and outlet N_2 flow rates, temperatures, and pressures were recorded. The study focused on the effects of N_2 flowrates and ambient server room temperatures on future cooling abilities for 180 min. The results of the experiment indicated that the flow speed influenced the cooling efficiency. More specifically, a flow rate of N2 at 50 kg/h cooled an artificial microclimate laboratory, 1.5 times more effectively, than a flow rate of 35 kg/h, with the former passive cooling that used a general renewable and natural cooling medium, at approximately 40 °C for 180 min and the latter cooling at approximately 44 °C over 180 min. Moreover, when the initiated room temperature was between 45 °C and 47 °C, cooling effects could be enhanced by 67% and as much as 1.08 times, when compared to a room temperature of 43 °C.

Fragkos et al. studied cutting-edge artificial intelligence in increasing server management by implementing heterogeneous multi-access edge computing (MEC) with specific MEC servers as inputs [25]. Bayesian Truth Serum (BTS) was used to validate collected information from users to minimize any avoidable error. Parameters used for experimentation included the size of the requested task, the computation resources required, required cache resources, the popularity of requested content, and mobility. The collected data from server users was used to compare and receive feedback on the effectiveness of the AI. It was found that the majority of users (73.33%) said "YES", and they were satisfied with the service provided by the MEC servers. Furthermore, it was concluded that an 80% prediction accuracy was more favorable than 90% and 70%, due to 80% having a higher system utility exceeding 2500, while 70% was approximately 2600 and 90% was approximately 2250. This outperformed the calculated mean prediction of 54.18%.

The application of neural networks to the prediction of data center network loads was explored through using of a multi-layer neural network, which received three inputs from an external information source [26]. The inputs were typically several requests on a cloud data server and how many times it was requested. Data were from the National Aeronautics and Space Administration (NASA) and the Environmental Protection Agency (EPA) WWW servers. It was concluded that a Linear Predictor model predicted the data more closely to the actual data than ANN and the RSME for the linear predictor was 0.25 while the ANN had an RSME of 8.00.

A study was conducted to explore neural networks' capability to manage and regulate services of power in data centers [27]. A feed-forward back propagation neural network was employed in the company of LSTM algorithms to predict resource and power consumption in approximately five minutes. The inputs to the neural network algorithm were the past and present CPU memory usage and power consumption and the output was the next time slot. Collected data was organized and reviewed using MATLAB R2020 coding with an output yielding a prediction with the intent of reducing costs. The results of the various trials indicated that the neural network combined with LSTM methods yielded a 20.6% cost reduction.

Cao et al. studied the implementation of neural networks into data centers to predict traffic by developing an algorithm called the Interactive Temporal Recurrent Convolution Network (ITRCN) in combination with architectures from CNN and GRU [28]. The study was designated to consider data from internet traffic such as email, messenger, news, music, and video as one-dimensional time series, and therefore the inputs to the ITRCN model. The outputs associated with the study were temporal and interactive features in the form of imaging, in which the CNN processes the inputs into an output vector format, and into temporal features for the GRU to learn from it. Information from email, games, news, videos, and web activity was used to experiment with different network data services. Through the trials, it was determined that the ITRCN was the most accurate prediction model of the experiment with an RMSE value of 1080. Also, the GRU model made more accurate predictions with an accuracy improvement of 13.1% in RMSE. The ITCRN model displayed an improvement of average accuracy by 13.6% in RSME when using “Yahoo!” data.

The role of neural networks in solving routing problems in data centers was explored by implementing an altered neural network traveling-salesman algorithm [29]. Inputs to the mathematical algorithm were values of bandwidth delay and communication link services. The neural network was compared with fully convolutional networks (FCN), recurrent networks (SimpleRNN, LSTM, GRU), and an RBF network, with outputs that measured optimal routes for the bandwidth-delay and metric of graph edges, ultimately to predict decisions for the network. Throughout the testing phase, it was found that a version of the FCN model was the most accurate, with 94.4% accuracy in training and 90.82% accuracy in testing. further analysis found that the FCN model predicted decisions accurately 91% of the time.

Yu et al. investigated the application of field-programmable gate arrays (FPGA) to improve existing and cutting-edge CNN [30]. The purpose of the study was to experiment and determine a more efficient solution to accelerators for applications within data centers. It was designated that cache weights and bandwidth information should be considered for inputs to the model, as the algorithm was planned to read and write bandwidth architecture as an output. Trials were conducted testing the acceleration of a variety of CNN models, such as FPGAs, and related enhancements elements such as enhanced processing element (EPE), supertile units (SU), and a 2D filter processing unit (FPU along with a server-provided graphics processing unit (GPU). The conclusion pointed out that the FPGA when compared to other models and programs, obtained the lowest latency of 3.8 ms. GPU achieved the highest FPS at 684 when measured at a batch size of 128, while FPGA performed exceptionally well with FPGA increasing the throughput of the servers by 149.2%, but with a 31.5% increase in costs.

Xiao et al. investigated the application of neural networking equalizers (NNE) for optimization the of computing observing through observing the NNE on transfer learning (TL) over different lengths [31]. The autoregressive recurrent neural networks (RNNE) model utilized fading nulls in the fiber dispersion and square law detections. The inputs to the model were past, present, and following samples of data. Outputs were the prediction accuracies of the RNNE. Through tests over the distances and convergence speeds of the transfers, it was determined that, over distances of 10km, 15km, 25km, and 30km the convergence speed increased by 23.81%, 14.28%, 9.52%, and 19.05% respectively using pruning algorithm assistance to the TL, based off of accurate predictions.

3.2. Managing Data Centers’ Cooling

Zhou et al. explored the utilization of DRL to optimize IT centers’ heat load using computation fluid dynamics (CFD) modeling [32]. They constructed a data-driven deep model using a LSTM neural network to predict the state of the data center in future instances, in terms of temperatures and power consumption. CFD model was developed to collect datasets useful for training that for unobtainable otherwise to validate DRL-based controller policies. The DRL models were applied to different and specific tasks such as load-aware cooling, IT optimization, and thermal-aware cooling. Compared to baseline data, the DRL consumed 15% less power for load-aware cooling, 10% for thermal-aware scheduling, and 15% for IT facility optimization.

Li et al. researched the competence of deep reinforcement learning to be used in optimizing green data centers and cooling systems within, using specifically the cooling control algorithm (CCA) and deep deterministic policy gradient (DDPG) [33]. Using deep reinforced learning algorithms and EnergyPlus, the overall aim and output of the study was to predict and observe the accuracy of a prediction model for energy costs for data centers, also with the ability to predict cooling loads. Inputs to the deep learning algorithm included temperatures of both the indirect and directive cooling temperatures, chilled water loop outlet temperature, and chilled colling air-loop outlet temperature. Information was used from historical data from the data center. Through experimentation, it was found that 11% of the cooling energy in comparison to the baseline could be reduced by prediction. The proposed model could also accurately predict control of the cooling while reducing the cooling energy costs by 15%.

Moore and Chase proposed a model to predict and manage cooling in data centers by a prototype machine learning automation software, “Weatherman,” which was developed by them [34]. The CFD model and the Fast Artificial Neural Net (FANN) library were used for designing the prototype. The prototype had sensitivity adjustments for accurate performance prediction. Input parameters included workloads, cooling settings, and room layout measurements. Cooling setting parameters are attributed to chilled air flow rates and air thermal properties, specifically temperature. It was concluded that a change to a singular, random parameter was almost negligible, while tweaking multiple parameters was what generated the most variance in accuracy. The results revealed that more than 75% of the predictions were within the 0.5 °C range and 92% of the predictions were within the 1.0 °C range.

The use of algorithmic incentives was investigated for scheduled data centers to prioritize green energy usage [34]. The experiment was built by four existing algorithms: Least Slack Time Rate First Round Robin (LSTRF-RR), Least Slack Time Round Robin (LST-RR), First Come First Served (FCFS), and Round Robin (RR). Considered inputs to these algorithms included temperature differences between the server room and threshold temperatures, electrical consumption for a server, the server threshold temperature, electrical consumption of the server during an increase in air conditioning, and power used to complete a task. The outputs for the study included power usage effectiveness (PUE) of the servers, saved cooling electricity, and fine incurred should the deadline be missed due to performance issues. An experiment designed to compare the selected algorithms by using job sets and workloads to compare electricity and PUE. The results showed that thermal aware LSTRF-RR saved approximately 18.05% of the electrical consumption, while others such as LST-RR, FCFS, and RR, saved 17.38%, 17.48%, and 17.56% respectively. After a week-long training, FCFS saved 17.56% of electricity consumption, while RR, LST-RR, and LSTRF-RR, respectively, saved 17.58%, 17.59%, and 17.25% of electrical consumption.

Chou and Bui studied the use of artificial intelligence and its applications on heating and cooling efficiency by comparing the performances of various artificial intelligence methods including SVR, ANN, classification and regression tree (CART), chi-squared automatic interaction detector (CHAID), and lastly general linear regression (GLR) [35]. Clementine (IBM SPSS Modeler) was used to experiment. Specific input parameters observed in the studied methods included relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. Output parameters included heating and cooling loads. The latter was measured in a range, with heating loads ranging from 6.01 kW to 43.1 kW with a mean of 22.31 kW, and cooling loads ranging from 10.9 kW to 48.03 kW, with a mean of 24.59 kW. The results indicated that compared, SVR and ANN had 25.2% to 63.3% lower, improved error rates than CL. SVR also had 48.1% to 88.8% lower, improved error rates than HL.

Wang et al. investigated the implementation of an ANN prediction algorithm into a data center to improve task scheduling with an overall goal to use the ANN to develop prediction models for thermal awareness to maintain a server’s optimal operation condition to improve and streamline response times [36]. A thermal-aware scheduling algorithm (TASA) was developed in which an internally located ANN predicted temperatures to communicate to the overall TASA system to mitigate overheating. For the basis of the study, an application of a homogeneous computer center was studied with inputs being the server room’s thermal properties, resource models, and incoming workloads, with an output of a prediction of the future thermal properties of the same server room. Experiments resulted in the TASA simulation being able to reduce the average maximum temperature by −14.1 °C for a simulated 30 days. Another finding was that for every −17.2 °C temperature reduction there was a 2% energy consumption saving. Ultimately, TASA saved up to 13.34% of the power consumption (5330 kW saved out of 80,000 kW) [36].

Chen et al. investigated temperature prediction in data centers using real-time data using a CFD (Fluent) simulation and real-time data-driven prediction algorithms [37]. Necessary data was collected by using more than 36 wireless, physical nodes that collect temperature data and air flow measurements. Inputs to the model included temperatures collected air flows, and data generated from training. The model was designed to measure runtime temperature and predict a temperature. For trials, five Groups were developed for controlling the CPU utilization, from idle state (2%) to full state (~90% utilization). The created groups were the layers of the server racks, Group 1 was towards the top, closest to the temperature sensor and ceiling vent that sucked the air out of the room, Group 2 was after, and amounting to Group 5, on the bottom of the rack closest to the AC inlet vent to the room. The simulations throughout the five groups showed that the average absolute prediction error was 0.3 °C. The RMSE values were less than 0.5 °C for most physical node locations, which further proved the accuracy of predictions. The RMSE values did not exceed 1.0 °C for the larger nodes. The error values on the outlets were slightly higher than those of the inlets (did not exceed 1 °C).

Matsuda et al. proposed a cloud-based automation system in server rooms, named the neural network the "Cloud-based Building Automation System" (CB-BAS) [38]. A combination of model predictive control (MPC) and Digital-Twin modeling was used to model and simulate the HVAC systems in server rooms. The input data includes temperatures of the ground, slabs, AC operation, humidity, electrical consumption of the batteries and lighting, and solar radiation with the intended output for an accurate prediction of optimal HVAC control parameters. Experimentation compared an ANN with a multiple linear regression (MLR) algorithm with both predicting HVAC behaviors regarding temperatures, humidity, and energy consumption. During testing, it was concluded that in the first 20 cycles, the average particle score measured around 0.12, decreasing as the cycle progressed. For the final 20 cycles, the score reached 0.10 which shows that the particles started to converge towards optimal parameters. The medians of the MAE measured approximately 0.05. The third quarter's MAE measured around 0.10, confirming accuracy. The ANN was then later compared to the MLR to observe accuracy readings and it was observed that the MAE of the ANN was 9.89 kWh and the MAE of the MLR was 11.18 kWh.

An adaptive learning-based thermal model (ALTM) was modeled to adjust temperature changes in server environments in collaboration with server room cooling systems using a neural network model [39]. The study aimed to develop a model to predict inlet temperatures based on the cooling system's operational parameters, such as server workload assignment. MATLAB toolkit was used for standard back-propagation methods and Levenberg-Marquardt algorithms in the training phase for the model. The validation of accuracy was set to 0.001 °C. The results of the experiment confirmed that the neural network accurately predicted the temperature of the inlet, based on the collected data. The average error in the neural network was calculated for 1.15 °C. By comparing a neural network to a CFD model, it was concluded that the CFD's average error was 2.1 °C. The neural network model accurately predicted the temperatures and adapted to changes in the room.

Gao, n.d. from Google investigated the application of artificial intelligence on data centers to optimize servers [40]. A three-layer neural network was implemented into the model by using Python programming software. Inputs to the model included IT load, number of chillers, weather conditions, and equipment setpoints. The model optimization of the PUE with a tolerance of 0.004+/0.005 was developed as a result with the calculated error of 0.4% for a PUE of 1.1. Also, it found that large efficiency gain occurs during the 0–70% load range. From 70% onwards, efficiency reached an asymptotic decline, at approximately a PUE of 1.07. It was concluded that when enthalpy increased in the simulation, the PUE value increased, too. At an enthalpy value of 1.0, the PUE reached 1.25.

A predictive model applicable to optimizing thermal efficiency in cooling systems in a data center was investigated by using a heat-flow reduced order model and a CFD simulation to verify models and create accurate thermally aware workloads [41]. The multi-objective particle swarm optimization (MOPSO) algorithm was used to optimize server workloads and decrease the overall existence of hot spots. The heat flow reduced order model (HFROM) was developed by MATLAB coding. Inputs to the models included air velocity, grid locations, air temperature, volumetric thermal expansion coefficient, and vortex strength. During experimentation, it was determined that the mean relative error (MRE), the standard deviation (STD) of the relative error, and correlation coefficients (CC) measured for maximum outlet temperature were 9.67, 5.61%, and 98.2% respectively. For the average server outlet temperature, the MRE, STD, and CC were 8.18%, 4.82%, and 91.2%, respectively. For the minimum server outlet temperature, the MRE, STD, and CC were 8.25%, 6.32%, and 82.6%, respectively. Applying the optimizing MOPSO algorithm, the coefficient of performance was increased by 17%, and workload allocation was increased by 10%.

3.2.1. Environment Monitoring and Prediction

AI-powered sensors continuously monitor temperature, humidity, and other environmental factors in the server room. By collecting real-time data, AI algorithms analyze the information to identify patterns, predict potential cooling issues, and optimize cooling strategies.

Damasio and Oyamada explored machine-learning techniques to develop a model to predict short-term cooling energy consumption using temperature and humidity levels in a server room with three machines [42]. Physical nodes collect relevant data using AZDelivery Digital's DHT22 temperature/humidity sensors. Inputs to the model included inside and outside humidity and temperatures collected from real instances of operation with an output that generated a command to inform the air condition system to change to a LOW, MEDIUM, or HIGH speed. During testing, initially, the air conditioning was set to 22 °C and studied for 106 days, with simple rules (SR) being studied for 21 and prediction rules (PR) being studied for 15 days. During this time, outdoor temperatures and time of day were used in analyzing the SR and PR. In the automatic mode, the consumption of energy was 44.71 kWh compared to the SR consumption of 36.94 kWh, a decrease of 17.37%. After another analysis for 87 h, the automatic mode consumed 72.22 kWh, and the PR system consumed only 44.89 kWh, with a decrease of 37.84%.

A team of researchers investigated humidity and temperature monitoring in a server room using a software called ZigBee 3.0 which connected a variety of hardware components, such as IR cameras, thermometers, microcontrollers, and computer components to communicate gathered data to the ZigBee algorithm to monitor the selected server room using thermal imaging [43]. Using ZigBee3.0, MATLAB 2020 coding, the program would trigger an alarm if the environmental conditions of the server room were not correct or hazardous for operation. It was concluded that the ZigBee 3.0 and MATLAB 2020 hybrid software could manage and communicate with various hardware components to observe humidity and temperature conditions. To validate the results, a two-tailed T-test was utilized to compare variations between measured and thermal imaging data. The results showed that measured data and thermal imaging ranged from 23.2 °C to 23.3 °C, with the thermal imaging having a high precision and the thermometer having a lower precision, generating truncation errors. The present error for humidity was 3.93788% which was acceptable and the confidence level of the experiment was calculated at 95% throughout the trials.

Asfahan et al. investigated the ability of artificial intelligence systems to predict the performance of cooling systems by studying the predictive cooling, humidity, and velocity demands on direct evaporative cooling (DEC), indirect evaporative cooling (IEC), and Maisotsenko evaporative cooling (MEC) [44]. An artificial neural network based on a deep learning framework and optimized Adam and RMSprop was developed. Inputs to the neural network included dry-bulb temperature, wet-bulb temperature, relative humidity, velocity, and area. Experimentation consisted of the analysis of the prediction accuracy using statistical error minimization models, with correlations between thermal properties, humidities, and variations in velocity being compared with each other. Using Adam and RMSprop, the ANN was able to reach accuracy levels at 98.6%, after a combined optimization. Mean average error values measured 4.44% for an outlet of the dry bulb temperature, 0.74% for humidity ratio, and 1.81% for outlet enthalpy.

Tuli et al. studied the use of AI techniques to create thermal profiles for data centers by developing an artificial intelligence algorithm for scheduling the Internet of Things (IoT) and Fog computing integration [45]. The Fog computing program was iThermFog and used to perform cutting-edge thermal-aware scheduling. iFogSim and ThermoSim as part of the Fog environment were used in the experiment and both software used CloudSim. Experiments were conducted comparing execution times with operation numbers between iThermoFog-Cloud, iThermoFog-Cloud + Fog, iThermoFog-Fog, game-theoretic thermal-aware allocation strategy (GTARA) GTARA-Cloud, and GTARA-Fog. Results determined that Fog was 14.44% to 18.92% faster in execution times than Cloud + Fog and Cloud, respectively. Furthermore, iThermoFog execution times were 52.68% and 9.35% faster than GTARA for Fog and Cloud, respectively. The energy consumption also decreased by 32% for Fog and 12% for Cloud programs. Also, Fog had lower temperatures respectively,10.55% and 13.46% in comparison to Cloud + Fog and Cloud.

Wang et al. collaborated to study a system that collected data to monitor thermal activity in a data center [46]. The CFD model was developed in the model-driven framework (MDF) and data-driven framework (DDF). The DDF used LSTM as its backbone while MDF was based on lumped-capacitance methods and ordinary differential equations. Inputs utilized for the study included server power, cooling power, and room temperature, with anticipated outputs measuring inlet and outlet air temperatures, and cooling enclosure. The MAPE and MSE were used to determine the accuracy of the models. The lumped-capacitance model was trained for over 20 h on datasets and then tested for predictability. The LSTM on the other hand was trained for less than 10 min. It was concluded that both models were accurate, while DDF was just slightly more predictive. The MAPE and MSE, respectively, were 3.36% and 0.47 for lumped capacitors. The MAPE and MSE, respectively, were 2.42% and 0.24, for LSTM.

Manaserh et al. investigated the capabilities of machine learning and its application to compact IT equipment, improved monitoring, and predictions of thermal and energy properties [47]. They utilized CFD software (6SigmaET R14) to create a prediction model for performance and combined the virtual simulation model with physical operational data. Inputs to the model included temperatures, power consumption, fan speed, CPU utilization, average temperatures, max CPU temperature, and CPU power consumption. The CFD model was developed as a turbulence model for the CPUs. The machine learning algorithms, random forest (RF), combined with the developed CFD model to create a dynamic CFD and it was extended by making a regression model, to develop a compact version. The physical models were compared to the virtual models to check the validity and compatibility of the compact models with other regression algorithms, such as linear regression, Lasso regression, linear SVM, and RBF. Results showed that the overall maximum mismatch for air velocity and air temperature from the general experimental model was 5.4% and 4.6%, respectively. Using a compact model, the team was able to generate data to compare to the general model. The maximum mismatch for the regression model's flow rate and outlet air temperature was 11.4% and 5.7%, respectively. The maximum error flow rates could be reduced from from155.4% to 11.4% and the air temperature error was also reduced from 30.2% to 5.7%.

The machine learning optimization for PUE of cooling of server rooms was studied by using SVM and Ant Colony Optimization (ACO) [48]. The input parameters to the model were room dimensions including height, equipment distance, air flow rate, inlet air temperature, and air leakage. The model was designed by using to predict energy and operational costs. The study aimed to observe machine learning optimization methods and to find through experiments how to use the machine learning techniques for optimization. The results showed that certain applications of machine learning, such as on-chip cooling and conventional airflow cooling methods, were optimized to cool more effectively. Conventional cooling was optimized to cool with a decrease in energy consumption of 38.9% and on-chip cooling with a decrease in power consumption at 60%.

The application of an ANN for optimizing algorithms for data centers was studied to develop an adaptive ANN model (AAM) and contribute control algorithm [49]. The input variables were server rack inlet temperature and cooling system energy, derived from air flow rate, heat transfer rate, and heat capacities of air to predict inlet temperature for server racks and cooling energy consumption. To ensure the ANN was suited for the experimentation of the datasets collected and used, 60% of the data was used to train the ANN. Another 10% of the datasets collected would be used to validate the ANN was capable of predictions, with another 10% being used for testing. The training, validation, and testing percentages were determined by use of the min-max normalization method. The results indicated that 99.88% of the time, the prediction algorithm was accurate, with an RMSE of 0.0598 °C.

The accuracy of predicting the temperature of server components based on past data by neural networks was investigated by employing immersion cooling through a two-part cooling system [50]. The ANNs that internally used non-linear functions, such as hyperbolic tangent (tanh), normalized exponential (SoftMax), and rectified linear (ReLu) were implemented. The ANNs RNNs, as the neural networks used the recurrent past data while the inputs to the RNN were the current temperature, past temperatures, and the current memory material. Outputs to the RNN ultimately would be a predicted future temperature, influenced by weighted parameter matrices. An array of experiments was conducted, and data were recorded to calculate the coefficient of determination and other accuracy calculations. Results revealed that the model was able to predict the temperature one minute before reaching it with an average error of 0.75 °C, with a root mean squared deviation (RMSD) of 0.957 °C, and an accuracy of 89.539%.

Chung et al. studied an ANN predicting model on energy consumption for cooling a server room by controlling the variable refrigerant flow (VRF) [51]. CFD model for simulating VRS of the data center was developed by coinciding neural network toolbox along with CFD software. It was designated that the temperature out, humidity out, temperature in, solar radiation, and load cooling as input parameters for the ANN algorithm to produce an output of a total energy consumption prediction, as well as temperatures and cooling loads. Results of different trials indicated that the temperature out, temperature in, and cooling load had a weak correlation with temperature out predicting 44%, temperature in predicting 35%, and load cooling predicting 29%. The CVRMSE for the tests was the lowest value of 15.49%.

Apte et al. applied the use of ThermoCast, a thermal forecasting model, to predict temperatures in server rooms to improve the predictability of the model [52]. CFD model of the server room components developed for thermodynamics and fluid mechanics studies. The inputs for the model were past dataset temperatures of components. The output parameters specified as prediction accuracies of ThermoCast compared to baseline information data. Simulation results indicated that the presented model allowed servers to operate at 13.75 °C or only 53 W for computing power and function while a server typically operated at 200 W for computing power, then the model operated at 26% computing power. The ThermoCast was shown to predict data at a higher percentage compared to the baseline at 71.4%. The mean ahead time (MAT) was also faster than the baseline, with the baseline being 2.3 min and ThermoCast being 4.2 min.

Song et al. studied the efficiency of computer room air conditioners (CRAC) to enable lower power consumption while providing adequate cooling for IT equipment [53]. The goal was to develop a dynamic model to predict and estimate temperatures for the data center and correlate it to the CRAC to continually provide adequate cooling. CFD simulations were used with turbulence models and temperature models within the simulation. Different cases were designed to test the model. Case 1 consisted of all server racks having their heating powers altered throughout the 3600s. Case 2 consisted of a realistic server rack heating power variance with the CRAC outlet temperatures set to a constant 24 °C. The generated output was delivered as an optimal system configuration based on temperature and energy-saving capabilities. Through experimentation, it was determined that for Case 1, the average power consumption was 20.66 kW. For Case 2, the temperature was set to 25 °C or 24 °C and integrated into an optimized mode, which alters energy consumption and other factors to increase performance based on trends and presented data. The optimal scenario for Case 2 saved 2.8% compared to the "self" mode of 3.1% at 24 °C.

Liu et al. explored a data-assisted first-principle model that used monitoring tools to predict server outlet air temperatures [54]. An intelligent platform management interface (IPMI) was employed for monitoring the server's parameters. The input variables included the temperature of the core of the CPU, the server's CPU usage, total server power, power of the CPU, and total power consumed by other server components to measure outputs for different scenarios. The experiment involved two primary scenarios, Scenario A consisted of input parameters temperature of the core of the CPU, power of the CPU, and total server fan speed to estimate server outlet temperature. Scenario B consisted of server inlet temperature and total fan speed to estimate the power of the CPU. Scenario B was used when the power of the CPU was unmeasurable, or the latency of predictions was lower than nine seconds. The experiment was conducted for 35 h with cold aisle temperatures varying from 25 °C to 37 °C, with a change in increments of 2 °C. When Scenario A and B were compared relative to their RSME, Scenario B had a slightly higher RSME in comparison with Scenario A over CPU usage levels measuring 10% to 90%. Scenario B's RSME values ranged from 0.03 to 0.14 higher in comparison to Scenario A. When Scenarios A and B were compared by latency predictions, model computation times were measured to be 8 milliseconds for Scenario A and 1.33 s for Scenario B. When total measured and recorded latency times were compared, Scenario A measured 9.01 s, and Scenario B measured 1.34 s.

The use of three-factor and four-factor orthogonal designs on a data center's thermal performance was studied through the applications of the orthogonal designs on under-floor air distribution and finding plenum geometric configurations using CFD software Airpak 3.0, Scenarios were described as E0–E7 for the trials conducted [55]. E0 consisted of a plenum height at 0.6m with a 20% perforation percentage, and an angle of detachable baffles that were non-existent, E1consisted of a plenum height at 0.6m with a 10% perforation percentage, and an angle of detachable baffles at 45°, E2 consisted of a plenum height at 0.6m with a perforation percentage at 20% and angle of detachable baffles at 60°, E3 consisted of a plenum height at 0.6m with 30% perforation percentage and angle of detachable baffles at 15°, E4 consisted of a plenum height at 0.6m with a 40% perforation percentage and angle of detachable baffles at 30°, E5 consisted of a plenum height that varied with an optimal combination, which also removed perforation percentage and angle of detachable baffles. Scenarios E0–E4 were conducted with an air temperature of 18 °C and an airflow rate of 3.5 m^3/s. E5–E7 were conducted with E5 having an air temperature of 16 °C, E6 having a temperature of 18 °C, and E7 having a temperature of 20 °C and having air flow rates of 2.9 m^3/s, 3.5 m^3/s, and 4.4 m^3/s respectively. scenario E0 having no plenum baffle the EAT was 31.3 °C and higher in scenario E1. During experimentation, it was observed that exhaust air temperature (EAT) was high in scenarios E0 and E1. It was noted that the EAT decreased as a trend for scenarios E2–E4. After experimenting with the perforation percentages and angles, while at a plenum height of 0.6 m, a perforation percentage of 20%, and an angle of around 60°, the optimal situation was determined. The return temperature index (RTI) and supply heat index (SHI) measured 1.04 and 0.63, respectively. The thermal performance evaluation index measured 0.62.

3.2.2. Adaptive Vent

AI optimizes the speed of cooling fans in response to varying server room conditions. By adjusting fan speeds based on real-time heat output, energy is conserved, and noise levels are minimized during periods of lower demand.

The adaptive vent tiles (AVT) to locally adjust and adapt to affect local and zonal cooling conditions was studied by using a MPC for thermodynamic control [56]. Components related to the computer room air conditioning (CRAC) that were studied included variable frequency drives and adaptive vent tiles. CRAC systems were responsible for the cooling in the studied server room. Inputs to the model included inlet air temperatures, fan speeds, and air mass flow rate. The integrated control of the variable frequency drives (VFD) and the AVTs were used in conducting experiments. During a five-hour trial, the VFD frequency changed by less than 10% and the average power consumption by the blowers was 5.1 kW and 36% less than in other trials conducted during this study. The supply air temperature (SAT) was integrated with the VFDs and AVTs while the threshold was set to 25 °C. It was observed that the violations of this threshold occasionally reached a maximum of 0.4 °C, with all vent tiles opened and maintained at 71% open during the experiment. It was noticed during the study, that CRAC unit information regarding power consumption was not available temporarily, but nominal total cooling power was observed. A decrease in using the MPC from 13.8 kW to 13.2 kW during the first hour was observed after the MPC had been enabled. Energy savings could also be attributed to the VFD set to 77.5% and SAT set to 18.3 °C.

Mirhoseini Nejad et al. studied machine learning in managing infrastructure for data centers was studied by a holistic data center infrastructure control (HDIC) framework [57]. The simple network management protocol (SNMP) was used to control the parameters, which included fan speeds and water flow rates within the heat exchangers in the cooling system. The neural network in the form of a nonlinear auto-regressive network with exogenous input was applied algorithm. It was found that HDIC saved 16% more power than the comparative cooling heterogeneity-aware infrastructure control (CHIC). Respectively, power consumption was 1574 W for CHIC and 1209 W for HDIC. The COP for CHIC was 3.7, while the COP for HDIC was 4.4.

Mousavi et al. investigated the adaptive automation architecture and the effectiveness of improving cooling in data centers by using a combination of Building Automation Systems (BAS) and Function Blocks to create an artificial intelligence network [58]. The created model for the server room contained local fans (LF) to cool CPUs locally and a Global Fan (GF) to pump in the outside air into the server room. The input data to the server room model was the entire server room energy consumption, CPU temperature, and air temperature in the server room. Data was collected from the experiment in increments of 900 s. In the first experiment, no fans were used. For the increments, at 900 s the CPU had a temperature of 85.8 °C with an energy consumption of 15.7 W. At 1800 s the CPU had a temperature of 114.5 °C with an energy consumption of 379.8 W. At the final increments at 2700 and 3600 s, the CPU had a temperature of 103.9 °C and energy consumption of 569.8 W and 133.0 °C with a power consumption of 824.7 W. In the second experiment the fans were utilized, and the CPU's temperatures measured 50.9 °C, 47.2 °C, and 46 °C, when the LFs were slow, medium, and fast, respectively. The energy consumption was 1803.3 W when the LFs were on slow, while fast speeds generated 2161.1 W. It was found that the temperatures of the CPUs did not reach the maximum value and therefore cooled effectively, and energy consumption was reduced by 18%.

The energy efficiency of smart cooling in server rooms was studied by thermal modeling and simulation [59]. Specifically, a cyber-physical holistic system (CPS) was modeled and used to optimize the cooling systems' energy efficiency. Parameters for the MATLAB model included power consumption in idle state and peak utilization, maximum rotation frequency, maximum power consumption of local fans, maximum and initial CPU temperature, specific air capacity, and the heat capacity of CPUs. During the experiment, by changing the fan speeds, the CPU's temperature was measured. A series of experiments were conducted for four scenarios. Each scenario was simulated for one hour, each taking place for 15 min, then switching to another scenario. Scenario 1 consisted of finding which fan speed was most efficient, Scenario 2 was set to optimize Scenario 1's findings by reducing operation time, Scenario 3 consisted of randomly toggling idle time for the cooling fans and finding a constant value, and Scenario 4 focused on energy reduction on the global fan by reducing operation time. In the first scenario, there was a saving of 17.9% (0.3931 kWh) of energy consumption, while Scenario 2 was scrapped by the team due to randomness in the data, and the third Scenario saved 18.1% (0.3979 kWh) of energy, and finally, Scenario 4 was triumphing in reducing the energy consumption by 39.8% (0.874 kWh).

Google researchers investigated the effects of predictive control on data center cooling using different predictive models [60]. The linear quadratic (LQ) control was employed as a certainty equivalence for updating the dynamic model throughout the process. A T-Markov model was developed as the base for the new simulation. Utilized inputs consisted of three categories: controls, states, and disturbances. Controls were manipulative variables such as fan speeds to control airflow and valve openings to regulate any water flow. States work to collect variables that were meant to be predicted and regulated such as differential air pressure and cold-aisle temperatures. Lastly, disturbances were conditions that were out of anyone's control and still affected the server racks, such as server power usage and inlet water temperatures. The experiment was conducted by using three models; Model 1 was trained 3 h with data and controls; this was the researchers' model; Model 2 was trained for a week on data generated by other controllers, and it was trained using 56 times more datasets than other models; Model 3 was trained for 3 h with data and controlled by a controller and it had a wider range of input data but did not have any exploratory actions. Results of the experiment indicated that Model 1 was the most cost-effective management system with 85.8% cost-effectiveness, while Models 2 and 3 were 77.2% and 87.9% respectively. When the three models were compared at "any" entering water temperature, Model 1's data was used 85.8% of the time, with Model 2's data being used 77.2% of the time, and Model 3's data being used 87.9% by controllers inside of the air handling units. Ultimately Model 1 was suggested, due to the lower cost.

The implementation of a neural network to predict transient temperatures using machine learning and thermofluidic transport equations was investigated by using ANNs with 3D zonal models, generated from a CFD simulation [61]. The gray-box was compared to a black-box model with the nonlinear autoregressive exogenous model (NARX). Inputs to the model included the cooling unit airflow rates, cooling unit set-point temperatures, and server workloads, as well as the velocity, and density of the server room air flows. The results of the testing indicated that using 4 hidden layers and 5 neurons yielded a training accuracy of 98% and a testing accuracy of 95%. Also, it was found that if cooling unit fan velocities decreased from 80% to 30% the temperatures dropped from 18 °C to 16 °C and the model continued to provide accurate predictions.

3.2.3. Mapping the Zones

AI algorithms can create thermal maps of the server room, identifying hotspots and areas with insufficient airflow. This information enables data center operators to reposition equipment and optimize airflow patterns for more effective cooling.

Hnayno et al. studied the capabilities of autonomously operated liquid-cooler power supply for data centers by using uninterrupted power supplies (UPS) in Northern France [62]. The goal was to provide a way for a way to cool servers in the event of a power loss. They created a simulated model of the UPS systems on SOLIDWORKS software, combining CFD and thermal simulations, and then tested the flow simulations, considering laminar and turbulent flows, modified with a k-ε model. Inputs to the model and simulations included temperatures, water flow rate, and internal and external ventilation. The physical model was fabricated to use thermal imaging for observing the variations of cold and hot regions. During experimentation, air-cooled and liquid-cooled UPS systems were compared. The Results showed that annually, the energy consumption for liquid-cooled was 14% less than the air-cooled. It was noticed that if the inlet water and ambient air temperatures decreased by 7–10 °C, the efficiency of the UPS increased by 8.2%, while the optimal flow rate was 20 L/min.

Grishina et al. investigated the utility impact of machine learning on thermal analysis for data centers using datasets and characteristics from an ENEA Portici CRESCO6 [62]. A clustering algorithm, k-means, as part of their machine learning components, was able to detect and group (cluster) server machines with similar surrounding temperatures. A period lasting from May-December 2018 and January-February 2019 was used to study thermal performance based on clustered server machines, in the ENEA-Portici Research Center (R.C.) using physical nodes to measure frequency ranges for thermal properties. It was noted that 50% to 86% of the physical nodes had the highest frequency in the medium temperature range (~37 °C), while 11% to 37% of the nodes were in the hot temperature range (41 °C), and 0.5% to 4% of the nodes were in the cold temperature range (30 °C).

A team of researchers investigated the capabilities of neural networks' prediction of airflow and temperature inside data centers [63]. The team used computational fluid dynamics and heat transfer (CFD/HT) simulations as part of the system to train the ANN, specifically Future Facilities 6Sigma Room. The proper orthogonal decomposition (POD) was employed to construct a framework while the ANN was trained using 300 CFD simulations. The error between the developed ANN and CFD simulations was less than 0.6 °C for the inlet temperature. The flow rate prediction error was calculated at 0.7%.

The optimization of airflow and temperature distribution in the data center by using a simulation-based ANN in combination function-based Genetic Algorithm (CFGA) was researched [64]. The CFD model of the server room for the study was developed to calculate Reynold's and Navier-Stoke's numbers, while the ANN was inputted with plenum height and floor tile percentage in an open area. The ANN was designed to control the building's HVAC systems. During the first optimization phase, the ANN yielded a higher air input temperature due to the input variables being 12.8 °C, a height of 0.3 m, and volumetric airflow of 4.8 m^3/s. During the second optimization, it was explored that if volumetric airflow increased to 6.6 m^3/s, the model yielded an error of less than 10%.

A team of researchers studied the use of a thermal supervisor that optimized airflow in data centers [65]. CFD model of the server room developed for simulation. The layout of the server room consisted of 27 server rack "tiles" that formed 4 rows, with row 3 missing a rack for other intended purposes. Inputs to the simulations include inflow properties and recirculation airflow, as well as volumetric flow rates for the air, which were in turn used to generate airflow models, used to measure volumetric flow rates of outlet air and control models to test the capacity of the tiles. Through simulations, it was concluded that for flow models, the average relative error measured below 1%. Specifically, it was found that flow rate surfaces at tiles 1, 7, 20, and 27 were kept constant at about 75% capacity, due to the nonlinear nature of the flow features under the plenum. These tiles were also located in rows with tiles 1 and 7 in the same row on the ends. Tiles 20 and 27 were also on the ends of rows 3 and 4.

Fang et al. used cutting-edge neural network implementation inside data centers to investigate the temperature distributions in a data center [66]. The goal of this study was to improve the older CFD models to map the flow rate between the selected server racks. A fast-temperature evaluation model (FTEM), backpropagation (BP), RBF, and extreme learning machine (ELM) architectures were used as the backbone for the neural network. Inputs to the neural network model included power, air flow rates, and reference temperatures. To test accuracy, usage of MAE, RMSE, and MAPE were designed to be calculated. Testing further showed that the accuracy of the FTEM was high, with a MAPE that measured less than 1%. It was shown that the FTEM-RBF combination with an absolute error of less than 0.4 °C was the more accurate method.

Cooling improvements in data centers were researched by using CFD modeling and simulations, physical models, and black-box DDM [67]. The goal was to develop a prediction model for row-based cooling to improve optimal temperature distribution and thermal performance. The Reynolds Averaged Navier-Stokes (RANS) models and turbulent models were used to simulate airflow for the CFD simulation. Inputs to the model included air density, air velocity, and time. Experimentation consisted of the developed model being compared to other algorithms, such as SVR, an artificial neural network, and Gaussian process regression (GPR). Compared to the other algorithms the proposed model generated a 4.2 °C increase of predicted temperature in the front chamber of server racks when a 4 °C increment in temperature was applied increasing the average error by 48%. It was also found that a 0.8 °C boost in average temperature was the result of an increase in 50% of the server workload and by changing the server locations from scattered to aggregated distribution, the hot areas changed, with 95% of the areas containing a prediction error of less than 1.5 °C.

A POD framework model was developed to predict temperatures in data centers [68]. The purpose was to develop a data-driven algorithm that worked in conjunction with the CRAC to provide energy-efficient cooling to server machines. Inputs to the model included the heat load, temperature, and time. Servers were categorized into different groups of servers, or zones for experimenting. The zones were labeled A–I, in a 3 × 3 grid pattern as shown in Figure 3. If the zone groups were to be on the numpad on a PC, A would be in the 5 positions, B at 8, C at 6, D at 9, E at 4, F at 7, G at 1, H at 2, and I at 3. For Zone A, the root mean square (RMS) deviation was 0.25 °C. It was found that the RMS values varied in a range of 0% to 6%, averaging 2.31% and the standard deviation averaging 1.66%. For Zone B, the RMS deviation was 0.6 °C, and no information was provided for the standard deviation or range of RMS values. For Zone C, the RMS deviation was 0.7 °C, and RMS ranges included values from 5% to 12%, with an average of 8.4%, and the standard deviation averaging 2.4%. For Zone D, the RMS deviation was 1.02 °C, and root mean square uncertainty was measured at 4.7%.

Figure 3. Testing Zones.

Optimizing the components in data centers to cut energy costs and consumption was studied, to observe the results related to the cooling system of microgrid implementation in data centers [69]. The adaptive vent tiles (AVT) integrated in conjunction with CRAH were modeled. Input to the study included cooling capacity, cooling costs, environmental conditions, exhaust temperature, and water supply temperature. For experimentation, horizontal and vertical integration of AVTs was evaluated. Vertical integration was conducted by use of real-time communication and coordination for the different cooling area distribution and delivery of cooled air for a zonal and local cooling ability, with a direct effect on the thermal properties of the server racks. Horizontal integration was conducted by the use of economically selecting energy to consume in terms of efficiency and cooling capacity. Further development led to the economizer used to determine costs was not solely capable of cooling the 330 kW IT load and needed further improvement by implementation of chiller-based mechanical cooling, with maximization on evaluating how much energy could be saved by swapping between the mechanical cooling and outside air cooling. Results showed the outside temperature fluctuated 10 °C for 24 h and despite this, the exhaust temperature measured a constant 30.5 °C. For this period, the air-side economizer had a cooling capacity of 3457 kWh, which led to an energy saving of 515 kWh in the model and 30% of annual energy.

3.2.4. Data-Driven Cooling

AI analyzes historical and real-time data to develop more efficient cooling strategies. For instance, it learns from patterns of server activity and environmental conditions to schedule cooling adjustments during peak and off-peak periods.

A Polish and Romanian collaborative research team studied the application of neural networks to predict energy expenditures and reuse waste heat generated by data centers [70]. The CFD models of server rooms were developed for thermodynamic calculations. The simulation based on buoyantBoussinesqPimpleFoam was conducted to capture buoyancy and pressure in the server rooms. The MLP neural networks predict thermal information. To ensure that the neural network was capable of use in the experiment, it was trained using 90% of the overall collected datasets. Furthermore, the neural network was tested using the remaining 10% of the dataset. Inputs to the model for running the simulation included cold air temperatures and airflow values as well as warm air generation within a server room. It was concluded that the model predicted an error rate of less than 0.1, which equates to less than 1 °C. Also, two options for providing savings and offering increased discount expenditure rates were established. The options were a power demand of 1.527 MW yearly, after reusing heat, and saving $ 110,971 with a payback time of 11.4 years, while the net present value has a 3% discount rate, equating to $64,642 over 15 years. The other option demanded 0 MW yearly, with a payback period of 5.7 years, costing $ 61,355 over 7 years.

4. Energy Consumption Optimization

AI facilitates optimizing the energy consumption of cooling systems by identifying opportunities for energy savings, such as adjusting cooling settings during non-peak hours or using free cooling techniques when ambient temperatures allow.

Shoukourian et al. investigated applications of machine learning for the cooling infrastructure of data centers in predicting power-saving capabilities [71]. Specifically, it was aimed to improve the COP of the cooling system. The model was based on using data gathered from the Leibniz Supercomputing Centre (LRZ) in Germany, which used a hot water-cooling system. The ANN was used to accommodate random drops in performance. Inputs for the ANN included the raw data, with a focus on the aggregated amount of cold produced by cooling circuits and power consumed by fans within the cooling towers, amount of cooling towers, in total and active, wet-bulb temperature, inlet water temperature and return water temperatures to each cooling circuit. Specifically, the LSTM network was the recurrent neural network used for initializing and using recurrent patterns and cycles. The neural networks were named IDA. Experimentation consisted of the ANN's accuracy in predicting COP by use of the mentioned inputs from the LRZ and its respective functioning thermal properties. Through experiments, it was found that the system and neural networks were capable of predicting irregularities in environmental conditions. It was also found that as the inlet temperature decreased, so did COP. When the temperature dropped from 39 °C to 30 °C, the cooling towers consumed more power and ultimately the COP decreased. The average error rate was estimated at 12.7%, but due to the randomness in performance, the team accepted this as a low error rate. Ultimately, the model was able to predict power-saving capabilities for 87.3%.

Liang et al. studied hybrid machine learning and its application to data centers by use of improving upon quality and operational costs [72]. CloudSim was used to compare different learning algorithms. A physical and virtual hybrid model was developed. A KNN algorithm as the type of machine learning algorithm was used to expand and learn using virtual machines. The hybrid model was given the name efficient hybrid machine learning (EHML) model, in which the KNN algorithm was expanded to cut down energy consumption and increase the performance of physical machines. FF, BF, EPS, and LAVMP algorithms were compared against the developed EHML algorithm through different trials. EHML performed superior by, consuming less energy around 10,000 kWh. EHML, during two different trials on average, utilized 91% and 77% of general-purpose virtual machines and hybrid virtual machines, respectively. EHML provided 29%, 26%, 17%, and 15% improved results relative to FF, BF, EPS, and LAVMP, respectively.

Patel et al. studied the capabilities of using LSTM learning networks to predict memory and CPU usage of virtual and physical machines, to prevent hotspots and coldspots within data centers [73]. An iOverbook model was used in which an ANN was employed by the method of two-layer feed-forward. Bidirectional LSTM (BiLSTM), and RNN were also used to process sequences over long-term dependencies. The trials consisted of an examination of performance using CloudSim when comparing LSTMs and BiLSTMs regarding resource management. The results of the prediction model revealed that the BiLSTM model used less storage than the LSTM on CPU usage, overall. The RMSE for the model fluctuated from 0.616 to 0.544 to 0.55 to 0.544 to 0.619. The proposed model consumed less energy as time advanced, with ~175 kWh being consumed at the 100% threshold and 310 kWh being consumed at the 30% threshold. Comparatively, the next best method consumed, physical machine (PM) Usage Prediction-aware virtual machine (VM) Consolidation (PUP-VMC), at ~320 kWh at a 30% threshold and 230 kWh at a 100% threshold.

Researchers examined the capabilities of reinforced learning for data centers geo-distributed to improve resource utilization and learning abilities based on historical data [74]. Markov decision process (MDP) was employed by way of differentiated Quality of Service (QoS) power consumption for large-scale problems, and CloudSim, using real datasets to generate models, then simulations, resource management, and allocations. Inputs to the models included power, QoS, and time. CloudSim simulations were employed to generate data centers and workloads. The results after nine days of the simulation were examined by comparing to dynamic voltage and frequency scaling (DVFS), static threshold (THR), median absolute deviation (MAD), interquartile range regression (IQR), and local regression (LR) methods algorithms. Results from the experiment revealed that the combined adaptive algorithms could increase QoS revenue over any other algorithms while consuming the least amount of power. It was found that the adaptive algorithm studied reduced power consumption by 13.3% in non-differentiated services and differentiated services by 9.6%.

Shaw et al. explored the use of reinforced learning to consolidate cloud data centers and improve policies that increase efficiency in dynamic environments, of VM, which were smaller independent machines that were part of a larger physical host [75]. To optimize the VMs Potential Based Reward Shaping (PBRS) was used to allow the device to increase learning speed and increase optimal decision-making. To further improve efficiency in methodology, the advanced reinforcement learning consolidation agent (ARLCA) was employed which was able to allocate an optimal amount of VMs to specific hosts, based on resource management maximization. Through experiments, ARLCA was compared to a cutting-edge PABFD framework by using CloudSim. The results of the ARLCA study resulted in a decrease in energy consumption by 25.35% with a confidence interval of 96% when compared to the PABFD framework. Over 30 days, the ARLCA reduced migrations by 49.17% with a confidence interval of 95%. Service-level agreement (SLA) violations fell by 63% during the modeling period.

Sharma and Garg investigated the application of an ANN to increase energy efficiency in data centers by dictating which resource should be given to a specific task, ultimately aiming to reduce energy consumption while improving performance [76]. The ANN was composed of the ANN-based Rack Scheduler (ARS) and the ANN-based Task Scheduler (ATS), with the ARS managing and deciding on rack-based task placement and the ATS then scheduled the tasks on available computers inside of the selected racks. Simulations of the models were performed on CloudSim and later organized using MATLAB coding, with inputs such as an incoming task and the current environmental state of the cloud. The ANN was given workloads and then went through training and validation modes before the trials could begin. Compared to the genetic algorithm (GA), MinMIN-MINMin (MM) algorithm, and linear regression (LR), trials could proceed. The results of using the ANN generated an accuracy of 99.9% based on the information provided and methods of backpropagation. In heavily loaded server environments, the ANN improves make span by 59%, decreases energy consumption by 45%, execution overhead by 88%, and active racks by 70%. Lightly loaded server environments, improved make span by 64%, decreased energy consumption by 71%, execution overhead by 43%, and active racks by 70%.

Berral et al. studied the utilization of machine learning in managing data centers to become power aware using an algorithm method based on the Best-Fit framework for a machine learning program [77]. OpenNebula and VirtualBox were employed to develop a model to study the performance in terms of energy, latencies, and data center environments. Parameters for the virtual models included resource usage, workloads, latencies, and sizes of the virtual machines. The team noted that consolidation of virtual machines could reduce power consumption by 30%, excluding energy savings on cooling overheads. The experiment of the mentioned virtual models and simulations regarding energy costs and energy consumption led to the conclusion to improve the energy consumption by 42% for consolidation and deconsolidation of virtual machines while keeping the SLA stable. For virtual machines that could not be consolidated, the benefits of the virtual machines improved by 2%. Prediction accuracy varied between virtual and physical machines, with virtual machines having a correlation ranging from 77.7% to 99.4% and physical machines correlating 90.9%.

Farahnakian et al. investigated the use of reinforcement learning to improve energy efficiency in the consolidation of virtual machines through a dynamic reinforced learning method, Q-learning, and a host power mode detector [78]. Q-learning used online learning by experiences from environmental factors and it was also equipped with a self-optimization capability by observing system states and choosing an action. Models and simulations were fabricated and executed using CloudSim, with the examined method of reinforcement learning-based dynamic consolidation (RL-DC). It was also compared to other methods such as LR, MAD, THR, and IQR. It was concluded that by use of the RL-DC algorithm, the data center energy consumption decreased by 12.5%, 19.4%, 22.6%, and 28.5%, compared to LR, MAD, THR, and IQR, respectively. Furthermore, using VM allocation techniques, RL-DC led to reducing power consumption by 15.6%, 33.6%, 44.3%, and 46.8% compared to LR, THR, MAD, and IQR methods, respectively.

Tarutani et al. studied the use of machine learning to cut power consumption for data centers utilizing the assistance of a regression model [79]. The team used the regression model on an augmented data center infrastructure management system (A-DCIM). Inputs to this prediction model included power consumption of servers, air intake consumption, power consumption of power supplies, current temperature distribution, operation parameters of outlet air temperature, and operation parameters of outlet air volume. The test of the model was based on the Keihanna data center (KDC) in Japan, in which there were a minimum of two rows of servers. After the experiment, the team concluded that the model was accurate to 0.095 °C and the machine learning system decreased power consumption by 30%.

Wang et al. examined the use of artificial intelligence to obtain more energy-efficient data centers using Blocking Island resource abstraction techniques along with other heuristic algorithms [80]. Inputs to the experimental AI system included the original network topology and a traffic matrix, which included the bandwidth requirements and reliability requirements. The study utilized algorithms such as the Elastic Tree heuristic algorithm, basic shortest-path algorithm, and power-aware heuristic mechanism, through varying trials. The experimentation consisted of evaluations of power savings compared with computation times, compared between power-aware heuristic mechanism (PHM), basic shortest-path (BSP) algorithm, and elastic tree, each variant compared with 1500 and 3000 instances. Findings showed that PHMalgorithms, performed faster, with 100% of the results generated occurring between 0.412 s and 0.921 s, compared to the 1.135 s and 3.142 s for Elastic Tree. Furthermore, different network loads were compared with power savings, experimentation between power-efficient network systems (PNS), using different fault tolerance levels, and the elastic tree algorithms occurred. The result confirmed that 30–50% power savings were feasible with PNS usage.

The use of ANNs to predict data center cooling strategies was studied by using a CFD model and 6SigmaRoom R14 to create the model to simulate temperatures and flow properties [81]. The Blackbox model of Information Technology Equipment (ITE) was the reference and DDM was employed to analyze data. With the use of the Blackbox model, parameters could be calculated such as air temperature, air pressure, airflow velocity, supply air temperature, and humidity. Output parameters affected by the inputs included the energy consumed by the heat exchangers. The neural network used Python to filter data along with NumPy and matplotlib for visualizing data. MATLAB was used for modeling, training, and confirming post-process procedures. Levenberg-Marquardt Algorithm (LMA) as part of ANN was used. It was determined the ANN was within an average of 3 kWh of error for consumed energy and airflow the model was within 0.2% error.

The implantation of an ANN into cooling systems and the optimization capabilities were studied by modeling indirect liquid cooling systems and the ANN's ability to predict the thermal conductance of components [82]. Inputs considered during the analysis of the ANN included thermal conductance, mass flow rates of cooling fluid, and outlet temperature of the cooling source. Experimentation consisted of the ANN being examined using actual data collected from heat loads of the server room. The experiments, primarily conducted on heat exchangers and variable frequency pumps (VFP) led to determining the maximum error presented for the heat loads for 6.40% which was acceptable and within the heat balance requirements. The ANN was also able to reduce the pump power consumption by 35.45% compared to previous experiments, further validating the model. The pumping power consumption was 27.34 W related to the optimal condition measured at a VFP frequency of 36.36 Hz.

The integration of data centers into smart grids and the energy-saving possibilities were studied by the direct response (DR) and the management using smart grids [83]. For this study, three methods were outlined, computing resource management, which included workload scheduling, geographical balancing of workloads, provisioning of servers, and dynamic voltage and frequency scaling, energy resource management, which included unit commitment and energy storage scheduling, and thermal load management, which included CRAC control and thermal storage system control. Inputs into the developed model were differences in electrical prices based on location if renewable energy was available to present further savings, energy resources available during the grid planning stage, computing resources, number of servers, computing requests, and load balancing of requests for the data centers. Results indicated that up to 44% of grid power dispatch was reduced by having dispatchable data centers incorporated. Also, it was concluded the income for the grid and cloud service providers could increase by at least 28% and 29%, respectively.

Kang et al. studied the role of artificial neural networks in controlling variable refrigerant flow (VRF) using the information provided by the Korean Meteorological Administration (KMA) for input variables in the data center [84]. Such variables included temperature out, humidity out, temperature in, cooling load, air temperature, condenser fluid temperature, and condenser fluid pressure. Output parameters from the study included cooling energy consumption prediction based on VFR operation inputs. The simulation was conducted using EnergyPlus and Building Controls Virtual Test Bed (BCVTB). The outcomes showed that the saved energy was 28.44% (95,509 kWh) while the prediction model was able to predict accurately with a CVRMSE of 10.3%

The thermal management of rack-level optimization in data centers was modeled by using an ANN to predict the total escaped power, capture index, and find the optimum alternative layout of equipment [85]. The selected predictive model used included an ANN model to achieve rapid temperature predictions and the use of CFD simulations to virtually construct CRAC components such as blowers and server fans. Inputs to the model included the total cooling power, heat loads, and thermal properties for the chilled water loop. The experiments were conducted over a time frame of 7.5 h, with data selected every 30 min. The genetic algorithm (GA) and constant cooling set-point were both implemented. Results indicated that the GA was the most optimal alternative solution, decreasing power usage and energy consumption by 20% compared to the constant cooling set-point. The GA solution was 52% more efficient at decreasing energy consumption than the constant cooling set-point with no return air temperature control.

The application of DRL to optimize energy efficiency in cooling systems for data centers was investigated by developing a DRL model-free architecture, Deep EE to optimize server job scheduling and adjustments of air flow rates in cooling systems [86]. Deep Q-Network (DQN) was implemented as a hybrid model without estimation or relaxation. Inputs into the system included power, thermal properties, and air flow rate. The test was conducted and parameterized action space DQN (PADQN), IT control optimization (ICO), cooling control optimization (CCO), joint IT and cooling optimization (JCO), and original DQN (O-DQN) algorithms were compared. The results indicated that PADQN saved more energy by 7%, 15%, 10%, and 5% when applied against ICO, CCO, JCO, and O-DQN, respectively.

5. Challenges

Based on the recent study, major barriers to AI-driven server room cooling include high initial capital costs for integrating AI with advanced hardware like sensors and adaptive systems, computational complexity requiring significant processing power that can strain data center resources, lack of standardized datasets for training AI models leading to inconsistent predictions, integration challenges with legacy HVAC infrastructure, scalability issues in handling extreme rack densities, real-time data latency risks during dynamic load fluctuations, and cybersecurity vulnerabilities from interconnected AI control systems. A particularly critical challenge involves:

(1) Sensor reliability, as these systems rely on real-time data from temperature, humidity, airflow, and pressure sensors for accurate predictions, yet sensor drift, calibration failures, or environmental degradation can trigger faulty inputs, cascading into inefficient cooling or overheating risks.
(2) Cybersecurity risks are equally pressing, with AI controls interconnected to building management networks vulnerable to breaches where attackers could manipulate cooling parameters, disrupt operations, or damage hardware through targeted algorithm attacks.
(3) Uncertainty in dynamic workloads, especially unpredictable heat spikes from AI/GPU-intensive tasks, complicates reliable forecasting, often requiring extensive model retraining or hybrid human oversight to maintain stability.
(4) Cooling challenges, Server rooms face significant cooling challenges due to high power density, uneven heat distribution, and the need for continuous, energy-efficient temperature control to prevent overheating and equipment failure.

The focus of this study was to explore the last challenge of the data centers.

6. Future Study Roadmap

The reliability and efficiency of data centers are closely tied to the effectiveness of their cooling management systems. As our world becomes increasingly digital, the integration of AI offers significant advantages in promoting more sustainable and energy-efficient data center operations. A review of existing literature reveals a variety of AI-driven approaches to temperature regulation within data centers, particularly through different implementations of ANNs. However, there are some drawbacks such as noisy data, model drift, and efficiency decline at high loads for ANN methods. Figure 4 depicts an AI-driven thermal management flowchart for data centers to map the interaction of the different components of the cooling system.

These studies generally, as shown in Figure 1, fall into three primary categories:

(1) Smart HVAC systems,
(2) Intelligent server room management (addressing both computational load and cooling), and
(3) Energy consumption optimization

A substantial portion of the research focuses on optimizing data center cooling using a range of strategies, including dynamic cooling control, environmental monitoring and forecasting, adaptive vent management, thermal zone mapping, and data-driven predictive cooling, as summarized in Figure 2, with some methodologies spanning multiple categories due to their interdisciplinary nature. Considering the well-established foundation of traditional HVAC systems, the development of smart HVAC systems emerges as the most promising path for future research, as integrating AI techniques into existing HVAC infrastructure can significantly enhance cooling performance while maintaining a relatively low level of implementation risk and leveraging the reliability of current technologies. At the same time, complementary approaches such as dynamic cooling, environmental monitoring and prediction, adaptive ventilation, zone mapping, and data-driven cooling focus more heavily on data analytics and AI-based management strategies, offering exciting opportunities to develop innovative methods for regulating and optimizing data center environments and transforming environmental control systems. Additionally, studies centered on energy consumption optimization in server rooms using AI techniques are strongly recommended, since their benefits extend beyond data centers to broader applications like district energy systems and other energy-intensive facilities, where reducing energy usage is a shared priority driven by both cost savings and the imperative to lessen environmental impact, making AI-enabled optimization a key area for future exploration and cross-sector implementation.

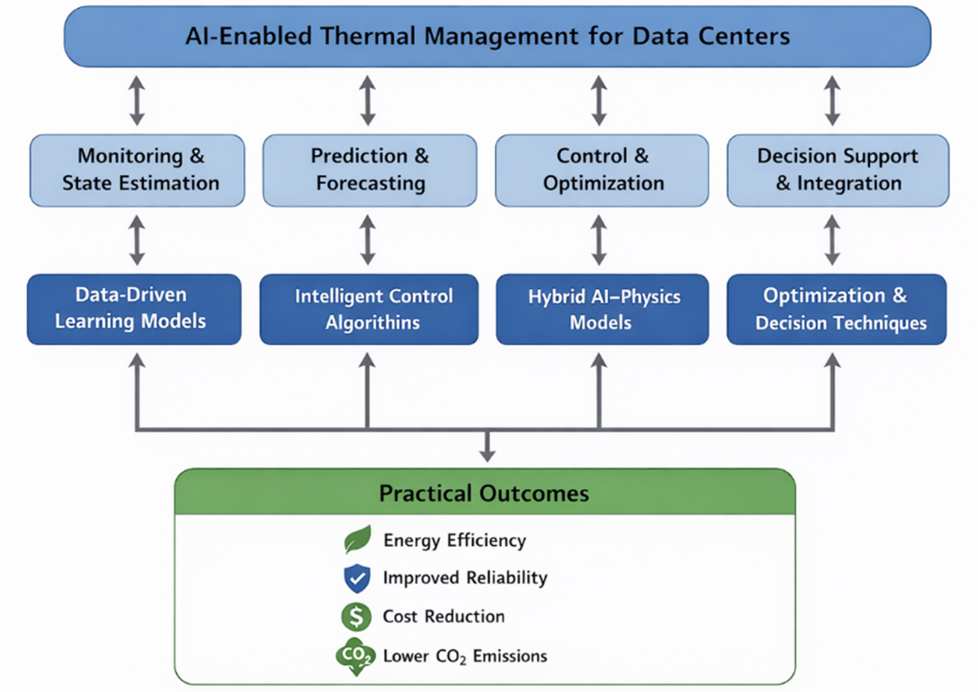

Figure 4. AI-driven thermal management flowchart for data centers.

Table 1 provides a structured overview of previous studies, classifying them based on their core strategies for managing and optimizing the data center environment.

Table 1. Categorizing past studies.

Smart HVAC	[3–17,61]
Dynamic Cooling Control	[4,5,9,10,12,34–44,46,54,63,71–74,76]
Environmental Monitoring & Prediction	[3,7,8,15,39–41,43–59,63,64,66,67,71,84,87]
Adaptive Vent	[13,59–64]
Mapping the Zone	[36,58,64–72]
Data-Driven Cooling	[9,10,34,40,45,72–84]
Energy Consumption Optimization	[3,6,9,11,13,37,41,45,49,51,52,54,57,59,62,72,74–89]

Future study opportunities in data center cooling, based on the provided paragraph, include:

Development of smart HVAC systems: by integrating AI techniques into existing HVAC infrastructure to enhance cooling performance with low implementation risk.

Data analytics and AI-based management strategies: focusing on dynamic cooling control, environmental monitoring and prediction, adaptive ventilation, thermal zone mapping, and data-driven predictive cooling.

Energy consumption optimization: using AI techniques, with applications extending beyond data centers to district energy systems and other energy-intensive facilities to achieve both cost savings and environmental sustainability goals.

Interdisciplinary research: combining thermal science, computer science, data science, electrical engineering, mechanical engineering, and material engineering. This collaborative approach can yield breakthrough innovations, such as advanced heat-conductive materials integrated with AI predictive algorithms and real-time thermal modeling.

Developing AI-based management strategies: to regulate and optimize the environment and distribution of load. These strategies can incorporate reinforcement learning for adaptive load balancing and multi-agent systems to coordinate cooling across distributed server clusters for maximum efficiency.

Future research opportunities are summarized in Figure 5.

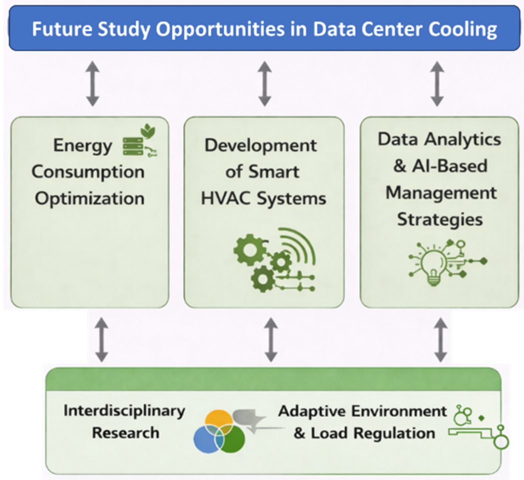

Figure 5. Future Study Roadmap on Data Center Cooling.

AI has clearly improved data center cooling, but several research gaps remain. Most existing methods, including reinforcement learning approaches, are validated mainly in simulation, with limited real-world deployment and safety validation on live facilities. Many models are developed by simulations, so there's growing interest in physics-informed or constrained learning that keeps predictions physically consistent and interpretable. As rack densities rise, most AI cooling research remains focused on air-cooled systems, leaving liquid and two-phase direct-to-chip cooling far less explored. There's also limited work on jointly optimizing cooling with IT workload scheduling and placement, rather than treating them separately, and models trained on one facility often fail to generalize to others due to differences in layout, climate, or hardware. Finally, the field lacks standardized benchmarks and long-horizon datasets for fair comparison, and as AI workloads dominate, efficiency metrics need to move beyond Power Usage Effectiveness (PUE) toward measures like power-to-compute performance.

7. Conclusions

The increasingly digital nature of modern society places a critical dependence on data centers and server rooms, making effective environmental control essential for reliable and continuous operation. Conventional HVAC systems, originally designed for comfort cooling, are often ill-suited to meet the dynamic and high-density thermal demands of data centers. In this context, Artificial Intelligence has emerged as a transformative tool, offering advanced capabilities in monitoring, analysis, and prediction that enable more precise control of thermal environments. AI-driven approaches have demonstrated significant potential to improve cooling efficiency, reduce energy consumption, extend hardware lifespan, and enhance overall system reliability. A comprehensive review of existing studies reveals three dominant research pathways in AI-based data center cooling: the development of smart HVAC systems, intelligent management of server load and cooling operations, and optimization of energy consumption. Prior research consistently shows that AI models outperform traditional methods in predicting temperature distribution and computational load, enabling proactive and adaptive control strategies. However, several challenges remain, including data availability and quality, model generalizability across different facilities and climates, integration with legacy HVAC infrastructure, real-time implementation constraints, and operational transparency for system operators.

Future research should focus on developing scalable AI-enabled HVAC systems with minimal implementation risk, advanced data analytics for dynamic cooling control and thermal zoning, and robust energy optimization frameworks applicable to both data centers and other energy-intensive facilities. Additionally, interdisciplinary collaboration across thermal science, computer science, and engineering disciplines is essential to advance AI-enhanced cooling technologies. Addressing these challenges through a structured research roadmap will be critical to realizing intelligent, energy-efficient, and resilient thermal management solutions for next-generation data centers.

Author Contributions

M.B.: data curation, writing—original draft preparation; B.R.: conceptualization, methodology, visualization, supervision, validation, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Harry Halloran Jr. and the American Refining Group Energy Institute at the University of Pittsburgh Bradford.

Institutional Review Board Statement

Not Applicable.

Informed Consent Statement

Not Applicable.

Data Availability Statement

Not Applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

AI has been used to improve the quality of design of graphics and for editing. Perplexity was used for improve the quality of designs of graphics. Claude was used for editing.

References

1. Montagud-Montalvá, C.; Navarro-Peris, E.; Gómez-Navarro, T.; et al. Recovery of waste heat from data centres for decarbonisation of university campuses in a Mediterranean climate. *Energy Convers. Manag.* **2023**, *290*, 117212. https://doi.org/10.1016/J.ENCONMAN.2023.117212.
2. Zhang, H.; Shao, S.; Xu, H.; et al. Integrated system of mechanical refrigeration and thermosyphon for free cooling of data centers. *Appl. Therm. Eng.* **2015**, *75*, 185–192. https://doi.org/10.1016/J.APPLTHERMALENG.2014.09.060.
3. Opalic, S.M.; Goodwin, M.; Jiao, L.; et al. ANN modelling of CO_2 refrigerant cooling system COP in a smart warehouse. *J. Clean. Prod.* **2020**, *260*, 120887. https://doi.org/10.1016/J.JCLEPRO.2020.120887.
4. Zhao, J.; Liu, X. A hybrid method of dynamic cooling and heating load forecasting for office buildings based on artificial intelligence and regression analysis. *Energy Build.* **2018**, *174*, 293–308. https://doi.org/10.1016/J.ENBUILD.2018.06.050.
5. Lee, D.; Ooka, R.; Matsuda, Y.; et al. Experimental analysis of artificial intelligence-based model predictive control for thermal energy storage under different cooling load conditions. *Sustain. Cities Soc.* **2022**, *79*, 103700. https://doi.org/10.1016/J.SCS.2022.103700.
6. Selvam, S., Kalaivanan, C., Jain, R.; et al. Artificial Intelligence Based Cooling System for Managing the Energy Efficiency Keywords: Recurrent Neural Network, Thermal efficiency, Smart buildings, HVAC system. *J. Contemp. Issues Bus. Gov.* **2021**, *27*, 1649–1659.
7. Hosoz, M.; Ertunc, H.M.; Bulgurcu, H. Performance prediction of a cooling tower using artificial neural network. *Energy Convers. Manag.* **2007**, *48*, 1349–1359. https://doi.org/10.1016/J.ENCONMAN.2006.06.024.
8. Yigit, K.S.; Ertunc, H.M. Prediction of the air temperature and humidity at the outlet of a cooling coil using neural networks. *Int. Commun. Heat Mass Transf.* **2006**, *33*, 898–907. https://doi.org/10.1016/J.ICHEATMASSTRANSFER.2006.04.003.
9. Li, A.; Xiao, F.; Zhang, C.; et al. Attention-based interpretable neural network for building cooling load prediction. *Appl. Energy* **2021**, *299*, 117238. https://doi.org/10.1016/j.apenergy.2021.117238.
10. Hou, Z.; Lian, Z.; Yao, Y.; et al. Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion technique. *Appl. Energy* **2006**, *83*, 1033–1046. https://doi.org/10.1016/J.APENERGY.2005.08.006.
11. Kamel, E.; Sheikh, S.; Huang, X. Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days. *Energy* **2020**, *206*, 118045. https://doi.org/10.1016/J.ENERGY.2020.118045.

12. Peng, Y.; Rysanek, A.; Nagy, Z.; et al. Using Machine Learning Techniques for Occupancy-prediction-based Cooling Control in Office Buildings. *Appl. Energy* **2018**, *211*, 1343–1358.
13. Sha, H.; Moujahed, M.; Qi, D. Machine learning-based cooling load prediction and optimal control for mechanical ventilative cooling in high-rise buildings. *Energy Build.* **2021**, *242*, 110980. https://doi.org/10.1016/j.enbuild.2021.110980.
14. Ma, Y. Model Predictive Control for the Operation of Building Cooling Systems. 2010. Available online: https://escholarship.org/uc/item/2hz3g57z (accessed on 6 June 2024).
15. Zabala, L.; Febres, J.; Sterling, R.; et al. Virtual testbed for model predictive control development in district cooling systems. *Renew. Sustain. Energy Rev.* **2020**, *129*, 109920. https://doi.org/10.1016/j.rser.2020.109920.
16. Xu, Z.Y.; Wang, R.Z. Simulation of solar cooling system based on variable effect LiBr-water absorption chiller. *Renew Energy* **2017**, *113*, 907–914. https://doi.org/10.1016/J.RENENE.2017.06.069.
17. Ramakrishnan, B.; Alissa, H.; Manousakis, I.; et al. CPU Overclocking: A Performance Assessment of Air, Cold Plates, and Two-Phase Immersion Cooling. *IEEE Trans. Compon. Packag. Manuf. Technol.* **2021**, *11*, 1703–1715. https://doi.org/10.1109/TCPMT.2021.3106026.
18. ACM Digital Library; Sigarch. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*; ACM: New York, NY, USA, 2009.
19. Liang, S.; Yang, Z.; Jin, F.; et al. Data Centers Job Scheduling with Deep Reinforcement Learning. In *Advances in Knowledge Discovery and Data Mining*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer: Cham, Switzerland, 2020; pp. 906–917. https://doi.org/10.1007/978-3-030-47436-2_68.
20. Wei, W.; Gu, H.; Wang, K.; et al. Multi-dimensional Resource Allocation in Distributed Data Centers using Deep Reinforcement Learning. *IEEE Trans. Netw. Serv. Manag.* **2022**, *20*, 1817–1829. https://doi.org/10.1109/TNSM.2022.3213575.
21. Gebreyesus, Y.; Dalton, D.; Nixon, S.; et al. Machine Learning for Data Center Optimizations: Feature Selection Using Shapley Additive exPlanation (SHAP). *Future Internet* **2023**, *15*, 88. https://doi.org/10.3390/fi15030088.
22. Fernández-Cerero, D.; Troyano, J.A.; Jakóbik, A.; et al. Machine learning regression to boost scheduling performance in hyper-scale cloud-computing data centres. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 3191–3203. https://doi.org/10.1016/J.JKSUCI.2022.04.008.
23. Nie, L.; Jiang, D.; Guo, L.; et al. Traffic Matrix Prediction and Estimation Based on Deep Learning for Data Center Networks. In Proceedings of the 2016 IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016.
24. Levy, M.; Raviv, D.; Baker, J. Data Center Predictions using MATLAB Machine Learning Toolbox. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 7–9 January 2019.
25. Guo, L.; Wu, J.; Hsieh, I.M.; et al. Advanced Single-phase Passive Immersion Cooling Solution with Natural Convection for Outdoor Edge Servers. In Proceedings of the 2021 20th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm), San Diego, CA, USA, 1–4 June 2021; pp. 417–422. https://doi.org/10.1109/ITherm51669.2021.9503238.
26. Wang, Y.; Zheng, W.; Shi, G.; et al. Experimental research on the influence of a novel passive cooling system on the data center temperature distribution. *Energy Build.* **2023**, *292*, 113135. https://doi.org/10.1016/J.ENBUILD.2023.113135.
27. Fragkos, G.; Lebien, S.; Tsiropoulou, E.E. Artificial intelligent multi-access edge computing servers management. *IEEE Access* **2020**, *8*, 171292–171304. https://doi.org/10.1109/ACCESS.2020.3025047.
28. Prevost, J.; Nagothu, K.; Kelley, B.; et al. Prediction of Cloud Data Center Networks Loads Using Stochastic and Neural Models. In Proceedings of the 2011 6th International Conference on System of Systems Engineering (SoSE), Albuquerque, NM, USA, 27–30 June 2011; pp. 276–281. https://doi.org/10.1109/SYSOSE.2011.5966610.
29. Liu, N.; Lin, X.; Wang, Y. Data center power management for regulation service using neural network-based power prediction. In Proceedings of the 2017 18th International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, USA, 14–15 March 2017; pp. 367–372. https://doi.org/10.1109/ISQED.2017.7918343.
30. Cao, X.; Zhong, Y.; Zhou, Y.; et al. Interactive temporal recurrent convolution network for traffic prediction in data centers. *IEEE Access* **2017**, *6*, 5276–5289. https://doi.org/10.1109/ACCESS.2017.2787696.
31. Perepelkin, D.; Ivanchikova, M. Research of Neural Network Architectures for Solving Adaptive Routing Problems in Multiprovider Networks of Distributed Data Centers. In Proceedings of the 2020 9th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 8–11 June 2020.
32. Yu, X.; Wang, Y.; Miao, J.; et al. A data-center FPGA acceleration platform for convolutional neural networks. In Proceedings of the 29th International Conference on Field-Programmable Logic and Applications (FPL 2019), Barcelona, Spain, 8–13 September 2019; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2019; pp. 151–158. https://doi.org/10.1109/FPL.2019.00032.
33. Xiao, J.; Sun, L.; Liu, C.; et al. Optimizations and investigations for transfer learning of iteratively pruned neural network equalizers for data center networking. *Opt. Express* **2022**, *30*, 36358. https://doi.org/10.1364/oe.468716.

34. Zhou, X.; Wang, R.; Wen, Y.; et al. Joint IT-Facility Optimization for Green Data Centers via Deep Reinforcement Learning. *IEEE Netw.* **2021**, *35*, 255–262. https://doi.org/10.1109/MNET.011.2100101.
35. Li, Y.; Wen, Y.; Tao, D.; et al. Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning. *IEEE Trans. Cybern.* **2020**, *50*, 2002–2013. https://doi.org/10.1109/TCYB.2019.2927410.
36. Moore, J.; Chase, J.S. Weatherman: Automated, Online, and Predictive Thermal Mapping and Management for Data Centers. In Proceedings of the 2006 IEEE International Conference on Autonomic Computing, Dublin, Ireland, 12–16 June 2006.
37. Haruna, A.A.; Jung, L.T.; Arputharaj, V.; et al. Incentive-Scheduling Algorithms to Provide Green Computational Data Center. *SN Comput. Sci.* **2021**, *2*, 252. https://doi.org/10.1007/s42979-021-00633-5.
38. Chou, J.S.; Bui, D.K. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy Build.* **2014**, *82*, 437–446. https://doi.org/10.1016/J.ENBUILD.2014.07.036.
39. Wang, L.; Von Laszewski, G.; Huang, F.; et al. Task scheduling with ANN-based temperature prediction in a data center: A simulation-based study. *Eng. Comput.* **2011**, *27*, 381–391. https://doi.org/10.1007/s00366-011-0211-4.
40. Chen, J.; Tan, R.; Wang, Y.; et al. A high-fidelity temperature distribution forecasting system for data centers. In Proceedings of the Real-Time Systems Symposium, San Juan, PR, USA, 4–7 December 2012; pp. 215–224. https://doi.org/10.1109/RTSS.2012.73.
41. Matsuda, Y.; Ooka, R.; Ikeda, S. The development of cloud-based building automation system and creating predictive models of HVAC system with machine learning. In Proceedings of the 2021 29th Mediterranean Conference on Control and Automation (MED), Puglia, Italy, 22–25 June 2021; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2021; pp. 955–960. https://doi.org/10.1109/MED51440.2021.9480357.
42. Mirhoseininejad, S.; García, F.M.; Badawy, G.; et al. ALTM: Adaptive learning-based thermal model for temperature predictions in data centers. In Proceedings of the 2019 IEEE Sustainability through ICT Summit (StICT), Montreal, QC, Canada, 18–19 June 2019.
43. Gao, J. Machine Learning Applications for Data Center Optimization. Available online: https://static.googleusercontent.com/media/research.google.com/zh-CN//pubs/archive/42542.pdf (accessed on 10 August 2026).
44. Khalaj, A.H.; Scherer, T.; Siriwardana, J.; et al. Multi-objective efficiency enhancement using workload spreading in an operational data center. *Appl. Energy* **2015**, *138*, 432–444. https://doi.org/10.1016/j.apenergy.2014.10.083.
45. Damasio, H.T.; Oyamada, M.S. A Predictive Actuator for IoT-based Air Conditioner Unit. In Proceedings of the 2020 X Brazilian Symposium on Computing Systems Engineering (SBESC), Florianopolis, Brazil, 24–27 November 2020.
46. Yumang, A.N.; Paglinawan, C.C.; Sejera, M.M.; et al. ZigBee Based Monitoring of Temperature and Humidity of Server Rooms using Thermal Imaging. In Proceedings of the 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2016), Penang, Malaysia, 25–27 November 2016; pp. 452–454.
47. Asfahan, H.M.; Sajjad, U.; Sultan, M.; et al. Artificial intelligence for the prediction of the thermal performance of evaporative cooling systems. *Energies* **2021**, *14*, 3946. https://doi.org/10.3390/en14133946.
48. Tuli, S.; Gill, S.; Casale, G.; et al. New trends and ideas iThermoFog: IoT-Fog based Automatic Thermal Profile Creation for Cloud Data Centers using Artificial Intelligence Techniques. *Internet Technol. Lett.* **2020**, *3*, e198.
49. Wang, Y.; Zhang, Y.; Nörtershäuser, D.; et al. Model and data driven transient thermal system modelings for contained data centers. *Energy Build.* **2022**, *258*, 111790. https://doi.org/10.1016/J.ENBUILD.2021.111790.
50. Manaserh, Y.M.; Tradat, M.I.; Bani-Hani, D.; et al. Machine learning assisted development of IT equipment compact models for data centers energy planning. *Appl. Energy* **2022**, *305*, 117846. https://doi.org/10.1016/j.apenergy.2021.117846.
51. Kumar, R.; Khatri, S.K.; Diván, M.J. Effect of Cooling Systems on the Energy Efficiency of Data Centers: Machine Learning Optimisation. In Proceedings of the 2020 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2–4 July 2020; pp. 596–600. https://doi.org/10.1109/COMPE49325.2020.9200088.
52. Choi, Y.J.; Park, B.R.; Hyun, J.Y.; et al. Development of an adaptive artificial neural network model and optimal control algorithm for a data center cyber–physical system. *Build. Environ.* **2022**, *210*, 108704. https://doi.org/10.1016/j.buildenv.2021.108704.
53. Pérez, J.; Pérez, S.; Moya, J.M.; et al. Thermal Prediction for Immersion Cooling Data Centers Based on Recurrent Neural Networks. In *Intelligent Data Engineering and Automated Learning—IDEAL 2018*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2018; pp. 491–498. https://doi.org/10.1007/978-3-030-03493-1_51.
54. Chung, M.H.; Yang, Y.K.; Lee, K.H.; et al. Application of artificial neural networks for determining energy-efficient operating set-points of the VRF cooling system. *Build. Environ.* **2017**, *125*, 77–87. https://doi.org/10.1016/J.BUILDENV.2017.08.044.
55. Li, L.; Liang, C.-J.M.; Liu, J.; et al. ThermoCast: A Cyber-Physical Forecasting Model for Data Centers. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1370–1378.

56. Song, M.; Chen, K.; Wang, J. Numerical Study on the Optimized Control of CRACs in a Data Center Based on a Fast Temperature-Predicting Model. *J. Energy Eng.* **2017**, *143*, 04017041. https://doi.org/10.1061/(ASCE)EY.1943-7897.0000460.
57. Liu, Y.; Van Le, D.; Tan, R. A data-assisted first-principle approach to modeling server outlet temperature in air free-cooled data centers. *Future Gener. Comput. Syst.* **2022**, *129*, 225–235. https://doi.org/10.1016/j.future.2021.12.003.
58. Chen, M.; Zhang, Z.; Deng, Q.; et al. Optimization of underfloor air distribution systems for data centers based on orthogonal test method: A case study. *Build. Environ.* **2023**, *232*, 110071. https://doi.org/10.1016/j.buildenv.2023.110071.
59. Zhou, R. et al., A Holistic and Optimal Approach for Data Center Cooling Management. In Proceedings of the 2011 American Control Conference, San Francisco, CA, USA, 29 June 2011–1 July 2011. https://doi.org/10.1109/ACC.2011.5991575.
60. MirhoseiniNejad, S.M.; Badawy, G.; Down, D.G. Holistic thermal-aware workload management and infrastructure control for heterogeneous data centers using machine learning. *Future Gener. Comput. Syst.* **2021**, *118*, 208–218. https://doi.org/10.1016/j.future.2021.01.007.
61. Mousavi, A.; Vyatkin, V.; Berezovskaya, Y.; et al. Cyber-physical Design of Data Centers Cooling Systems Automation. In Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 20–22 August 2015.
62. Lazic, N.; Boutilier, C.; Lu, T.; et al. Data Center Cooling Using Model-Predictive Control. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montreal, QC, Canada, 2–8 December 2018; pp. 3818–3827.
63. Asgari, S.; MirhoseiniNejad, S.M.; Moazamigoodarzi, H.; et al. A gray-box model for real-time transient temperature predictions in data centers. *Appl. Therm. Eng.* **2021**, *185*, 116319. https://doi.org/10.1016/j.applthermaleng.2020.116319.
64. Hnayno, M.; Chehade, A.; Klaba, H.; et al. Experimental investigation of an autonomous liquid-cooled uninterruptible power supply (UPS). *Energy Rep.* **2023**, *9*, 5775–5787. https://doi.org/10.1016/J.EGYR.2023.05.010.
65. Grishina, A.; Chinnici, M.; Kor, A.L.; et al. A machine learning solution for data center thermal characteristics analysis. *Energies* **2020**, *13*, 4378. https://doi.org/10.3390/en13174378.
66. Athavale, J.; Yoda, M.; Joshi, Y. Artificial Neural Network Based Prediction of Temperature and Flow Profile in Data Centers. In Proceedings of the 2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm 2018), San Diego, CA, USA, 29 May–1 June 2018; pp. 871–880. https://doi.org/10.1109/ITHERM.2018.8419607
67. Song, Z.; Murray, B.T.; Sammakia, B. Airflow and temperature distribution optimization in data centers using artificial neural networks. *Int. J. Heat Mass Transf.* **2013**, *64*, 80–90. https://doi.org/10.1016/J.IJHEATMASSTRANSFER.2013.04.017.
68. Lucchese, R.; Johansson, A. Coldspot: A Thermal Supervisor Aimed at Server Rooms Implementing a Raised Plenum Cooling Setup. In Proceedings of the 2019 American Control Conference (ACC), Philadelphia, PA, USA, 10–12 July 2019; pp. 5870–5877.
69. Fang, Q.; Li, Z.; Wang, Y.; et al. A neural-network enhanced modeling method for real-time evaluation of the temperature distribution in a data center. *Neural Comput. Appl.* **2019**, *31*, 8379–8391. https://doi.org/10.1007/s00521-019-04508-y.
70. Asgari, S.; Moazamigoodarzi, H.; Tsai, P.J.; et al. Hybrid Surrogate Model for Online Temperature and Pressure Predictions in Data Centers. *Future Gener. Comput. Syst.* **2021**, *114*, 531–547.
71. Ghosh, R.; Joshi, Y. Rapid temperature predictions in data centers using multi-parameter proper orthogonal decomposition. *Numer. Heat Transf. Part A Appl.* **2014**, *66*, 41–63. https://doi.org/10.1080/10407782.2013.869090.
72. Zhou, R.; Wang, Z.; Mcreynolds, A.; et al. Optimization and Control of Cooling Microgrids for Data Centers. In Proceedings of the 13th InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, San Diego, CA, USA, 30 May–1 June 2012.
73. Antal, M.; Cioara, T.; Anghel, I.; et al. Reuse of data center waste heat in nearby neighborhoods: A neural networks-based prediction model. *Energies* **2019**, *12*, 814. https://doi.org/10.3390/en12050814.
74. Shoukourian, H.; Wilde, T.; Labrenz, D.; et al. Using machine learning for data center cooling infrastructure efficiency prediction. In Proceedings of the 2017 IEEE 31st International Parallel and Distributed Processing Symposium Workshops (IPDPSW 2017), Orlando, FL, USA, 29 May–2 June 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; pp. 954–963. https://doi.org/10.1109/IPDPSW.2017.25.
75. Liang, B.; Wu, D.; Wu, P.; et al. An energy-aware resource deployment algorithm for cloud data centers based on dynamic hybrid machine learning. *Knowl. Based Syst.* **2021**, *222*, 107020. https://doi.org/10.1016/j.knosys.2021.107020.
76. Patel, Y.S.; Jaiswal, R.; Misra, R. Deep learning-based multivariate resource utilization prediction for hotspots and coldspots mitigation in green cloud data centers. *J. Supercomput.* **2022**, *78*, 5806–5855. https://doi.org/10.1007/s11227-021-04107-6.
77. Zhou, X.; Wang, K.; Jia, W.; et al. Reinforcement Learning-Based Adaptive Resource Management of Differentiated Services in Geo-Distributed Data Centers. In Proceedings of the 2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS), Vilanova i la Geltrú, Spain, 14–16 June 2017; pp. 1–6.

78. Shaw, R.; Howley, E.; Barrett, E. An Advanced Reinforcement Learning Approach for Energy-Aware Virtual Machine Consolidation in Cloud Data Centers. In Proceedings of the 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), Cambridge, UK, 11–14 December 2017.
79. Sharma, M.; Garg, R. An artificial neural network based approach for energy efficient task scheduling in cloud data centers. *Sustain. Comput. Inform. Syst.* **2020**, *26*, 100373. https://doi.org/10.1016/j.suscom.2020.100373.
80. Berral, J.L.; Gavaldà, R.; Torres, J. Power-Aware Multi-Data Center management using machine learning. In Proceedings of the 2013 42nd International Conference on Parallel Processing, Lyon, France, 1–4 October 2013; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2013; pp. 858–867. https://doi.org/10.1109/ICPP.2013.102.
81. Farahnakian, F.; Liljeberg, P.; Plosila, J. Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning. In Proceedings of the 2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2014), Turin, Italy, 12–14 February 2014; pp. 500–507. https://doi.org/10.1109/PDP.2014.109.
82. Tarutani, Y. et al. Temperature distribution prediction in data centers for decreasing power consumption by machine learning. In Proceedings of the IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom 2015), Vancouver, BC, Canada, 30 November–3 December 2015; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2016; pp. 635–642. https://doi.org/10.1109/CloudCom.2015.49.
83. Wang, T.; Xia, Y.; Muppala, J.; et al. Achieving Energy Efficiency in Data Centers Using an Artificial Intelligence Abstraction Model. *IEEE Trans. Cloud Comput.* **2015**, *6*, 612–624. https://doi.org/10.1109/tcc.2015.2511720.
84. Simon, V.S.; Siddarth, A.; Agonafer, D. Artificial Neural Network Based Prediction of Control Strategies for Multiple Air-Cooling Units in a Raised-floor Data Center. In Proceedings of the 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), Orlando, FL, USA, 21–23 July 2020.
85. Shao, W.; Chen, Q.; He, K.; et al. Operation Optimization of Liquid Cooling Systems in Data Centers by the Heat Current Method and Artificial Neural Network. *J. Therm. Sci.* **2020**, *29*, 1063–1075. https://doi.org/10.1007/s11630-020-1283-5.
86. Guo, C.; Luo, F.; Cai, Z.; et al. Integrated energy systems of data centers and smart grids: State-of-the-art and future opportunities. *Appl. Energy* **2021**, *301*, 117474. https://doi.org/10.1016/J.APENERGY.2021.117474.
87. Kang, I.; Lee, K.H.; Lee, J.H.; et al. Artificial neural network-based control of a variable refrigerant flow system in the cooling season. *Energies* **2018**, *11*, 1643. https://doiorg/10.3390/en11071643.
88. Athavale, J.; Yoda, M.; Joshi, Y. Genetic algorithm based cooling energy optimization of data centers. *Int. J. Numer. Methods Heat Fluid Flow* **2020**, *31*, 3148–3168. https://doi.org/10.1108/HFF-01-2020-0036.
89. Ran, Y.; Hu, H.; Zhou, X.; et al. DeepEE: Joint optimization of job scheduling and cooling control for data center energy efficiency using deep reinforcement learning. In Proceedings of the International Conference on Distributed Computing Systems, Dallas, TX, USA, 7–10 July 2019; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2019, pp. 645–655. https://doi.org/10.1109/ICDCS.2019.00070.