Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

Machine Learning for Improved Root Cause Analysis of Data Center Energy Inefficiency

By:

Elsa Abreha

Advisor:

Dr. -Ing. Dereje Hailemariam

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master Science in Telecommunication Engineering.

Jun 18, 2025
Addis Ababa,Ethiopia

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering
Telecommunication Engineering Graduate Program

Machine Learning for Improved Root Cause Analysis of Data Center Energy Inefficiency

By:

Elsa Abreha

Members of the examining committee's names and signatures:

Dr. -Ing. Dereje Hailemariam

Advisor	**Signature**	**Date**
Internal Examiner	**Signature**	**Date**
External Examiner	**Signature**	**Date**
Chair or School Dean	**Signature**	**Date**

Declaration

I, Elsa Abreha, hereby declare that this thesis titled 'Machine learning for improved root cause analysis of data center energy inefficiency' is my original work. It has been conducted under the guidance and supervision of Dr. -Ing. Dereje Hailemariam. All materials, data and information sourced from external references have been appropriately cited and acknowledged in accordance with academic standards.

Elsa Abreha _____
Name Signature

This thesis has been submitted for examination with my approval as a university advisor.

Dr. -Ing. Dereje Hallemariam _____
Advisor Signature

Abstract

In this study the vital issue of energy inefficiency in the data center is addressed by developing a machine learning-based framework for root cause analysis (RCA) of high power use efficiency (PUE) rates. Focusing on the Nefas Silk Data Center, the research leverages a 1D Convolutional Neural Network (CNN) model to classify PUE efficiency and employs SHapley Additive exPlanations (SHAP) to interpret the contributions of key operational features. The dataset, comprising 6,586 hourly measurements, identifies air conditioning systems as the primary driver of inefficiency, followed by UPS losses during power conversion and distribution, and rectifier performances.

The suggested 1D CNN model demonstrates outstanding performance, achieving an accuracy of 99.99%, sensitivity of 99.99%, and an F1 score of 99.99%, outperforming the comparative LSTM and RNN architectures. By integrating global and local interpretability methods, the framework provides recommendations to optimize energy consumption, reduce operational costs, and improve sustainability. The findings underscore the potential of machine learning to transform data center energy management, offering a scalable solution to improve efficiency in similar infrastructures. Future work will aim to improve energy efficiency in data centers by improving root cause analysis through the integration of historical data from all data centers at the core site. It will also involve comparative studies to assess regional factors that influence performance. These initiatives seek to create a robust framework for sustainable energy management in various environments.

Keywords: Energy efficiency in the data center, Power Usage Effectiveness (PUE), Root Cause Analysis (RCA), 1D CNN, SHAP values, Machine Learning Interpretability.

Acknowledgment

I want to begin by expressing my deep gratitude to Almighty God for giving me the strength, courage, and guidance throughout this thesis journey. I am especially grateful to my advisor, Dr.-Ing. Dereje Hailemariam, whose unwavering support, patience, and motivation have been the cornerstone of my research. I also thank ethio telecom for providing the sponsorship that made this work possible.

A special thanks goes to Dr. Eng. Yihenew Wondie and Dr. Yalemzewd Negash for their insightful comments and constructive feedback during my thesis presentations. Your suggestions truly enriched my work.

Lastly, I am extremely grateful to my loving husband and my children. Your constant support, unconditional love, and understanding have allowed me to complete this investigation.

Dedication

This thesis is dedicated to my beloved family. To my husband, whose unwavering support and encouragement have been my greatest strength; and to my children, who inspire me every day with their love and joy. Your patience and understanding during this journey have meant the world to me. I also dedicate this work to all those who have supported me along the way, believing in my dreams and aspirations.

Table of Contents

List of Figures

List of Tables

List of Acronyms

1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
3P	Three-Phase
AC	Alternating Current
ACO	Ant Colony Optimization
AdaGrad	Adaptive Gradient
AI	Artificial intelligence
ANN	Artificial Neural Networks
ASHRAE	American Society of Heating, Refrigerating and Air Conditioning Engineers
ATS	Automatic Transfer Switches
CDNN	Convolutional Dense Neural Network
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
CRAC	Computer Room Air Conditioning
DC	Direct Current
DCIE	Data Center Infrastructure Efficiency
DL	Deep Learning
DNN	Deep Neural Networks
EP&EMS	ElasticNet Power and Environment Monitoring System
GPU	Graphics Processing Units
GSA	Global Sensitivity Analysis
HDDs	Hard Disk Drives
IoT	Internet of Things
IP	Internet Protocol

IQR	Interquartile Range
IT	Information Technology
KW	Kilowatt
LightGBM	Light Gradient-Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long-Short-Term Memory
MAC	Media Access Control
ML	Machine Learning
PDB	Power Distribution Board
PDF	Power Distribution Frames
PUE	Power Usage Effectiveness
RCA	Root Cause Analysis
ReLU	Rectified Linear Unit
RF	Random Forest
RH	Relative Humidity
RNN	Re- current Neural Networks
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
UPS	Uninterruptible Power supplies
VRLA	Valve Regulated Lead Acid Batteries
XGBoost	Extreme Gradient Boosting

1 Introduction

This chapter introduces this thesis, which aims to enhance energy efficiency performance in data center using a machine learning-based root cause analysis (RCA) approach. It reviews the existing literature, identifies research gaps, and outlines the scope and limitations of the work. The chapter also presents the thesis contributions to the field, emphasizes the innovative aspects of the methodology, and provides an overview of the thesis structure.

1.1 Background

Data centers are critical infrastructures in modern information ecosystems that serve as the backbone of data processing, storage, and transmission [1][2]. As the volume of data generated worldwide continues to grow exponentially, the energy demand of data centers has risen, necessitating innovative strategies to optimize energy use while managing increasing workloads. Efficiently designed data centers host critical information technology (IT) equipment and supporting infrastructure, which makes them essential for organizations that rely on IT services to operate effectively [3, 4].

As the global data center infrastructure grows rapidly, energy consumption becomes more concentrated, which requires an energy efficient to become an essential priority for sustainable operation. Data centers consume approximately **1-2%** of the global electricity demand, a figure projected to increase as digital services expand [15, 16]. This trend of growth in energy consumption requires the immediate development of innovative methods to reduce energy use while increasing operational efficiency.

The architecture of data centers comprises two main components: *critical operational systems* and *supporting infrastructure* [11]. Critical operational systems include IT equipment such as servers, storage, and network equipment, which are usually arranged in rows of racks for efficient management. The supporting infrastructure includes electrical systems, uninterruptible power supplies (UPS), rectifiers, power distribution frames (PDF), and mechanical systems designed to maintain optimal operating conditions for IT equipment [12]. These systems rely on high voltage electricity supplied by utility companies,

which is transformed into lower voltages suitable for IT use.

However, the unstable nature of the power supply requires the use of backup systems, such as diesel generators, to ensure continuous operation. During power outages, automatic transfer switches (ATS) activate these backup systems, while UPS systems provide immediate power until full operational capacity is restored [13]. This redundancy, while essential for reliability, contributes to the significant energy consumption of data centers, leading to increased operational costs and environmental concerns, particularly in terms of carbon emissions [4].

To evaluate data centers' energy efficiency of, the Power Usage Effectiveness (PUE) metric is widely used. PUE is expressed as the ratio of the total power entering a data center to the power utilized by the IT equipment. A PUE value closer to **1** indicates that a larger proportion of the energy is being used directly for computing tasks, reflecting a higher operational efficiency [14]. However, many data centers experience inefficiencies due to complex operational conditions, such as inefficiencies in the cooling system, power distribution losses, and fluctuating IT workloads [15, 13]. Such inefficiencies underscore the need for advanced analytical methods to accurately identify and address the root causes of energy waste.

To ensure optimal performance and reliability in traditional (core sites) data center operations, tools like the ElasticNet Power and Environment Monitoring system (EP&EMS) provide real-time monitoring and management. This tool continuously tracks critical energy and environmental metrics, such as power consumption, temperature, humidity, and airflow, offering operators valuable insight into the operational environment. Using real-time data analysis, the EP&EMS tool identifies anomalies and deviations from normal parameters, generating customizable alerts that facilitate proactive problem resolution and minimize downtime. However, even though these monitoring tools work well in detecting issues, they often fail to pinpoint the underlying causes of inefficiencies.

Traditional methods for analyzing energy inefficiency in data centers often rely on manual diagnostics or rule-based systems. Although these approaches have provided valuable information, they are limited by their inability to capture intricate, nonlinear relationships among various factors that influence PUE. For example, cooling inefficiencies, power distribution losses, and fluctuating IT workloads interact in complex ways that traditional methods cannot adequately address [12, 17]. This limitation underscores

the need for advanced analytical methods capable of uncovering hidden patterns and providing actionable insights for Root Cause Analysis (RCA).

To overcome the limitations of traditional diagnostic approaches in RCA, machine learning (ML) has emerged as a powerful tool to address complex energy management problems in data centers [13]. Among ML techniques, Convolutional Neural Networks (CNN), which are a type of Deep Learning (DL), are especially effective in extracting patterns from time series and spatial data [18, 19]. When combined with SHapley Additive exPlanations (SHAP), these methods provide a robust framework for analyzing the root causes of energy inefficiencies. SHAP improves the interpretability of machine learning models by explaining the contribution of individual features to the predictions of the models,enabling data center operators to make informed, data-driven decisions[24].

This thesis suggests a novel framework that integrates CNN and SHAP to perform RCA of energy inefficiencies in the data center, with a specific focus on improving PUE. CNN is used to detect temporal and spatial patterns in operational data, while SHAP explains the contribution of individual features to PUE fluctuations, enabling the identification of key factors driving inefficiencies. Together, these techniques offer a comprehensive approach to understanding the underlying causes of high PUE values, allowing operators to implement targeted strategies to improve energy efficiency [23, 26].

The framework is validated using real-world data from the Nefas Silk Data Center, incorporating metrics such as cooling system performance, power system utilization, and environmental conditions. The results demonstrate the potential of ML-driven RCA frameworks to support energy savings strategies, reduce operational costs, and advance sustainability goals. These findings offer practical value for data center operators seeking data-informed approaches to improve energy efficiency.

1.2 Statement of the Problem

ethio telecom operates 35 traditional data centers at its core site in Addis Ababa and other regions, many of which face significant challenges related to energy inefficiency and lack of adherence to green energy strategies. The main problem motivating this study is illustrated in Figure 1, which shows a random selection of 25 data centers. The data, collected from the EP&EMS over a two-month period, reveal considerable inconsistencies

between the energy consumed by (IT) equipment and the total power supplied. These inconsistencies indicate significant inefficiencies in energy use. When total input power spikes without a corresponding increase in data center load, substantial energy waste is suggested,resulting in high PUE rates. This inefficiency not only indicates under-utilization of resources, but also leads to increased operational costs due to unnecessary energy expenditures. Furthermore, excess energy consumption contributes to a larger carbon footprint, raising environmental concerns and underscoring the critical need for better energy management techniques to boost sustainability and operational effectiveness.

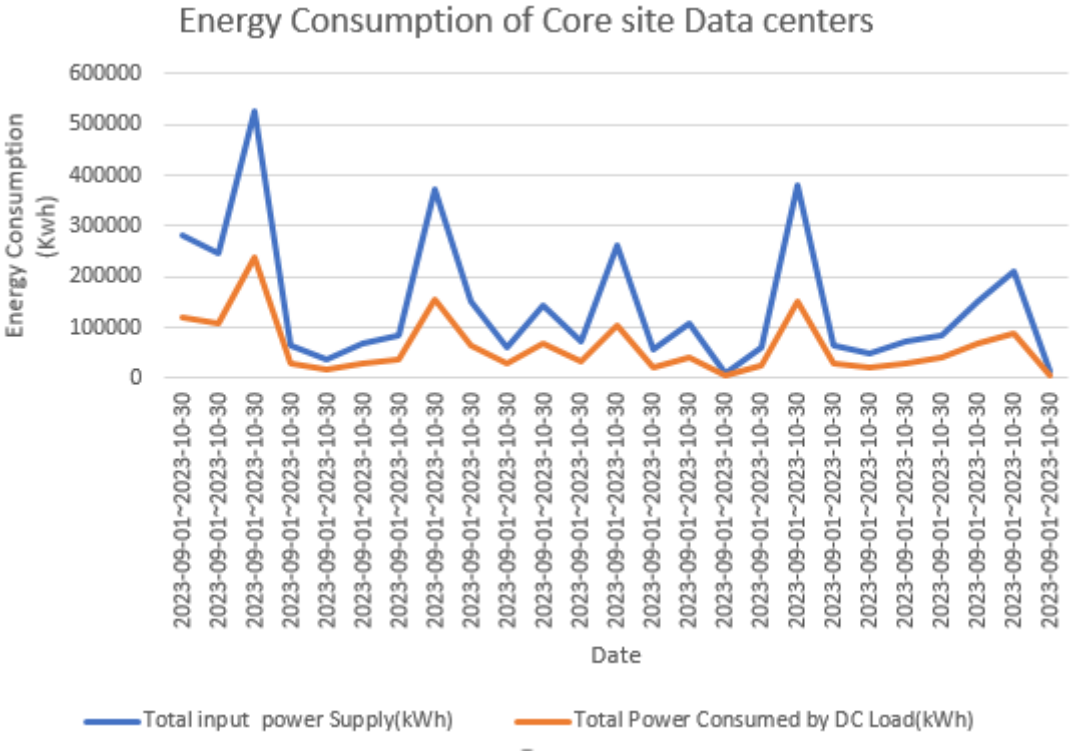

Figure 1: Power consumption in core site data centers source: [Elastic monitoring tool].

Data center performance can be inefficient due to various factors, including cooling system inefficiencies, under-use of resources, inefficient power distribution, legacy equipment, and improper load balancing. For example, outdated cooling systems can consume excessive energy without effectively maintaining optimal temperatures for IT equipment [25]. Furthermore, underutilized servers can lead to a disproportionate energy consumption compared to actual computing workload [27]. If operators do not identify, assess and anticipate the root causes of PUE inefficiency, it can result in substantial energy losses. Therefore, it is essential for operators to understand these causes and implement correc-

tive measures promptly. In doing so, they can optimize operations proactively, prevent future inefficiencies, and improve the overall energy performance of data centers [4].

This thesis aims to investigate the following research questions:

1. What are the key factors contributing to energy inefficiency in these data centers?
2. What are the root causes of the energy discrepancies observed between IT equipment consumption and total power supply?
3. How can ML techniques be applied to analyze energy consumption patterns and uncover the underlying drivers of inefficiency in traditional data centers?
4. What actionable recommendations can be proposed to improve energy efficiency and reduce waste in these data centers?

Previous research on data center energy inefficiency has used Global Sensitivity Analysis (GSA), a traditional sensitivity analysis technique that has limitations due to its assumption of linear relationships between input variables and outputs and its dependence on analytical models [11]. In contrast, recent advances in ML have expanded the capabilities of RCA by enabling more robust and data-driven identification of causal factors, especially in systems characterized by nonlinear interactions and large datasets. Therefore, there is a critical need to investigate RCA methods that use ML algorithms to uncover the root causes of high PUE and persistent energy inefficiencies in these data centers.

1.3 Objectives

1.3.1 General Objective

The main objective of this research is to develope a ML-bases RCA method to identify energy inefficiencies in the data center. This will involve utilizing DL techniques integrate with SHAP to improve the interpretability and understanding of the predictions of the model.

1.3.2 Specific Objectives

To achieve the general objective, the study pursues the following specific objectives:

- Examine data center infrastructure and performance metrics by analyzing design and operational characteristics, with a focus on metrics that significantly impact energy efficiency.
- Conduct a comprehensive review of the literature to assess existing research on energy efficiency of data centers and identify relevant methodologies and findings.
- Identify key factors that influence energy efficiency in data center operations.
- Collect and compile relevant historical data that reflect the impact of identified factors on energy performance.
- Prepare and preprocess the data using appropriate techniques such as cleaning, normalization, and formatting to ensure analytical readiness.
- Develop,evaluate and compare a DL model: Create a DL classification model specifically designed to predict energy inefficiencies in the data center.
- Integrate SHAP with the DL model to assess the contributions of different input features to high PUE rates.
- Interpret the results of the model, draw conclusions, and provide actionable recommendations based on the analysis.

1.4 Literature Review

Zerihun et al. [11] investigate the root causes of energy inefficiency at the ethio telecom Legehar data center, focusing on the factors that contribute to increasing energy consumption. They employ a combination of ML techniques and Sobol-Global Sensitivity Analysis (GSA) to analyze operational data and identify key factors that contribute to inefficiency. Using Random Forest (RF) for feature selection and GSA to quantify the impact of selected features on energy efficiency, the study identifies UPS efficiency and cooling system performance as the primary contributors to the data center's high PUE value of 2.34. This research highlights the importance of addressing power and cooling inefficiencies to improve data center performance.

Marcinkevics et al. [25] propose a novel approach to improve energy efficiency in existing data centers by addressing ineffective cooling implementations. The authors suggest using IoT enabled ultrasonic airflow and temperature sensors to measure air circulation and thermal conditions in server rooms. The data collected are analyzed to create a 3D model of the facility's air management, identifying areas of high bypass airflow and recirculation that reduce energy efficiency. PUE is used as a key metric to evaluate energy efficiency and categorize efficiency levels according to PUE values. For small data center loads, the authors recommend using variable-flow Computer Room Air Conditioning (CRAC) units, while for high-density loads, containment methods such as cold-aisle, hot-aisle, or rack exhaust containment are suggested. However, the study acknowledges the limitations in scaling the approach to large data centers and its generalizability across diverse operational environments.

Baskoro et al. [28] explore the application of advanced machine learning techniques to enhance prediction accuracy and reduce downtime in data center operations by identifying root causes of performance issues. The study involves collecting extensive datasets, pre-processing for consistency, and developing a Convolutional Dense Neural Network (CDNN) architecture that combines convolutional layers for feature extraction with dense layers for classification. Although the study demonstrates the effectiveness of DL in operational efficiency, it also highlights limitations such as the dependence on high-quality data and the need for extensive training to achieve optimal model performance. This research underscores the potential of DL in data center management, while

acknowledging challenges related to implementation.

Leka et al. [29] address the growing need for efficient workload management in cloud environments by developing a robust forecasting model that combines CNN and LSTM networks. The methodology involves data pre-processing, feature extraction using CNN to capture spatial patterns, and LSTM for modeling temporal dependencies in workload data. The hybrid model demonstrates high accuracy in predicting virtual machine workloads, enabling better resource allocation and energy efficiency. However, the study notes limitations such as the model's sensitivity to hyperparameter tuning and the requirement for extensive historical data. The findings emphasize the benefits of integrating complementary neural network architectures to adapt to dynamic workload patterns and enhance data center efficiency.

Sathupadi [27] investigates the challenges of managing cloud clusters in data centers, focusing on improving scalability and operational efficiency in the context of increasing workloads. The study develops a deep learning framework that classifies cloud cluster performance and optimizes resource allocation. Using algorithms such as CNN and RNN, the framework analyzes performance metrics such as central processing unit (CPU), memory usage, and network traffic. The classification model distinguishes between optimal and suboptimal cluster configurations, enabling proactive resource management. However, the study highlights challenges such as the dependence on labeled data for training and the complexity of integrating the framework into existing management systems. The research underscores the potential of deep learning to enhance scalability and performance in data center operations.

Gebreyesus et al. [33] explore the use of machine learning techniques, specifically SHAP, to optimize data center operations by improving feature selection processes in predictive models. The methodology involves applying SHAP to determine the most influence features that influence key performance metrics and improve the interpretability and efficiency of the model. The study uses regression and tree based models to analyze how operational characteristics affect data center performance. Although the approach shows significant improvements in model accuracy and interpretability, the authors recognize challenges such as computational overhead and the risk of overfitting if feature selection is not carefully managed. The study emphasizes the importance of the relevance of the features in improving both the performance of the model and the decision-making

in data center management.

The reviewed literature highlights the growing use of ML and DL techniques to improve energy efficiency in data centers. However, most studies focus on prediction and resource optimization, with limited attention to interpretable RCA of energy inefficiencies. Few have explored integrating deep learning models with explainability tools such as SHAP to trace the underlying drivers of high PUE. This study addresses this gap by proposing a CNN–SHAP-based RCA framework tailored for traditional data centers.

1.5 Methodology

Develop, evaluate and compare comprehensive research methodology focused on exploring RCA of energy inefficiency in the data center. The methodology is structured into five key phases: literature review, data collection, data pre-processing, model selection using a DL model, and evaluation. Each phase is designed to systematically address the challenges associated with high PUE rates and provide actionable insights to improve energy efficiency.

The initial step involved conducting an extensive literature review to gain a thorough understanding of current studies related to high PUE rates and RCA. This review established a solid foundation of knowledge, allowing for the identification of key themes and gaps within existing research. The literature review focused on advanced analytical techniques, particularly machine learning, to analyze the complex relationships that contribute to energy inefficiencies in the data center.

Data were collected using the EP&EMS tool, a platform designed to monitor and analyze the energy consumption of the data center. Historical data was gathered from this tool for the ethio telecom core site, providing a comprehensive data set necessary to identify the potential causes of high PUE rates. The data set includes metrics such as power consumption, temperature, humidity, and airflow, all of which are essential for understanding the dynamics of energy consumption and environmental conditions within the data center.

Once the data was collected, data pre-processing techniques were applied to ensure its

quality and suitability for analysis. This involved cleaning the data by addressing missing values, removing outliers, and normalizing the features to maintain consistency. In addition, feature selection was performed to create new variables that capture important relationships in the data. These preprocessing steps were vital to ensure the integrity of the data and prepare it for analysis using machine learning algorithms.

The next phase involved model selection using a DL model to establish causal relationships between various factors and issues of energy inefficiency. The DL model was chosen for its ability to capture complex and nonlinear relationships in the data,which makes it well suited to identify the most influential factors contributing to energy inefficiency. The DL architecture consists of multiple layers, including input, hidden, and output layers, with activation functions such as ReLU to introduce non-linearity. The model was trained using preprocessed data and its hyperparameters, such as the number of layers, neurons, and the learning rate, were carefully tuned to optimize performance. This approach enables the model to make accurate predictions about instances of PUE inefficiency.

The performance and effectiveness of the selected DL model were rigorously evaluated using metrics such as accuracy, precision, sensitivity / recall and F1 score. These metrics provide a comprehensive assessment of the model's ability to classify energy inefficiencies and identify root causes. Furthermore, an analysis of feature significance was carried out utilizing SHAP to identify the most significant factors contributing to energy inefficiency. SHAP values provide interpretable insights into the contributions of individual features to the model's predictions, enabling a nuanced understanding of the underlying factors driving energy inefficiency in data center operations. This combined approach of using a DL model and SHAP for RCA guides the development of targeted improvement strategies.

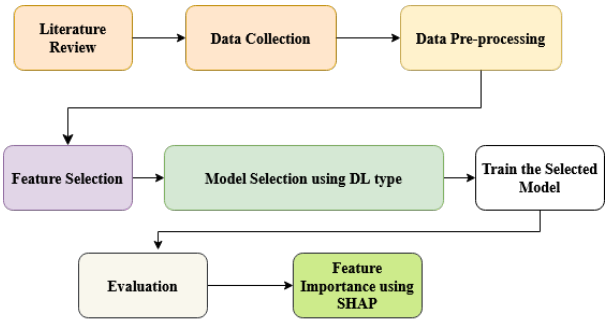

Figure 2: Methodology

1.6 Scope and Limitations

1.6.1 Scope

The scope of this research focuses on designing a RCA to identify the key factors that contribute to the high PUE rate or energy inefficiency at the Nefas Silk Data Center in Addis Ababa. The study uses a trained DL classification model to predict energy inefficiency and incorporates SHAP values to assess the influence of various input features on the PUE rates.

1.6.2 Limitations

This study data collection challenges due to restricted access to the EP&EMS. Although the research successfully mapped the flow of power consumption from the main distribution board to IT loads, direct access to real-time PUE measurements and quantitative granular datasets was limited. To address this gap, PUE values were manually derived using available data and established computational methods that model the dynamics of the power flow within the data center.

1.7 Contributions

This thesis contributes an innovative methodology to analyze energy inefficiency in the data center, particularly with regard to high PUE rates. It establishes a robust

framework that integrates DL with SHAP to perform advanced examinations of historical data from the EP&EMS tool. This integration improves predictive accuracy and provides clear insight into the variables that affect inefficiency, allowing the identification of the root causes of high PUE. This research offers practical recommendations for data center operators to reduce energy waste and improve efficiency. Identifying root causes provides targeted strategies to reduce PUE, reduce operational costs, and promote sustainable energy practices in data centers. Ultimately, the research highlights the effectiveness of advanced analytical techniques in understanding the complex energy dynamics of data centers, advocating for more sustainable operational practices in energy performance management.

1.8 Thesis Structure

This thesis is structured to systematically explore energy inefficiency in the data center, focusing on high PUE rates. Chapter 2 provides an overview of operations, components, and the concept of root cause analysis RCA. Chapter 3 discusses the machine learning techniques used, including DL and SHAP, explaining their relevance to the analysis of energy inefficiency. Chapter 4 introduces the 1D-CNN model designed for RCA, detailing its development and integration with SHAP for interpretability. Chapter 5 presents the results, evaluating the performance of the model in identifying the root causes of inefficiency. Finally, Chapter 6 concludes the study, summarizing key findings, contributions, and future research directions, emphasizing the importance of data-driven approaches to improving energy efficiency and sustainability in data centers.

2 Data Center Operations, Energy Efficiency, and Root Cause Analysis

This chapter explores the essential concepts and operational principles of data centers, focusing on design considerations, classifications by tier, and operational aspects. In addition, it examines energy management practices and discusses key metrics used to evaluate energy efficiency and overall performance.

2.1 Fundamentals of Data Center Operations

Data centers in the telecommunications sector are vital infrastructures that support the storage, processing and transmission of large amounts of data. They enable essential services such as Internet connectivity, cloud computing, mobile networks, and real-time communication platforms [1]. These infrastructures are built to manage massive amounts of data traffic with minimal latency, ensuring seamless connectivity for users. Data center operations are complex and require careful planning and management to maintain high availability, scalability, and energy efficiency [30]. As the demand for data-driven services continues to grow, the role of data centers in telecommunications has become more critical, making their efficient operation a top priority [31].

One of the core principles of data center operations is high availability, must operate 24/7 to ensure uninterrupted service, as even a few minutes of downtime can result in significant financial losses and damage to the reputation of an organization [1]. To achieve this, data centers rely on redundant systems, including backup power supplies such as uninterruptible power supplies (UPS) and diesel generators, as well as failover mechanisms to prevent service interruptions [16]. These systems ensure that data centers can continue to function even during power outages or equipment failures, maintaining the reliability of telecommunication services[14].

Given their substantial electricity consumption, energy efficiency has become a fundamental concern in data center operations, as these facilities contribute significantly to global energy demand. Inefficient energy use not only increases operational costs but also contributes to environmental challenges, such as higher carbon emissions [32]. To address this, data centers employ various strategies to optimize energy consumption, such as improving cooling efficiency, adopting energy-efficient tools, and using renewable energy sources [37]. Metrics such as PUE are used to evaluate energy efficiency, with lower PUE values indicating better performance [34].

Scalability is also a key consideration in data center operations. As data demands grow, data centers' infrastructure has to be scalable in order to handle increasing workloads [35]. This requires flexible designs, modular hardware, and the ability to integrate new technologies without disrupting existing operations. Scalability ensures that data centers can meet the growing needs of the telecommunications sector, supporting the expansion of services such as fifth generation (5G) networks and edge computing.

Finally, security is a critical component of data center operations. Telecommunication data centers store sensitive information, making them prime targets for cyberattacks and physical breaches [36]. Robust security measures, including firewalls, encryption, access controls, and surveillance systems, are essential to protect data and ensure compliance with regulatory requirements. By addressing these fundamentals—high availability, energy efficiency, scalability, and security—data centers in the telecommunications sector can provide reliable and efficient services while minimizing operational costs and environmental impact [10].

2.2 Data Center Tier Classification

Data centers are classified into tiers based on their redundancy, availability, and overall reliability of the infrastructure. The Uptime Institute, a globally recognized organization, has established a tier classification system that ranges from Tier I to Tier IV. Each level represents a different level of performance, fault tolerance, and redundancy, with Tier IV being the most advanced and reliable. This classification system helps organizations evaluate and select data centers that meet their specific needs for uptime, resilience, and

operational continuity [30] [38] [39] [40].

2.2.1 Tier I: Basic capacity

Tier I data centers are the most basic and least complex type, designed for small businesses or organizations with minimal IT requirements. They have a single path for power and cooling distribution and lack redundant components, making them suitable for non-critical applications where occasional downtime is acceptable. The key characteristics of Tier I data centers are summarized as follows.

- **Uptime:** 99.671% (28.8 hours of annual downtime).
- **Redundancy:** No redundancy in power or cooling systems.
- **Fault Tolerance:** No fault tolerance; Any failure in the infrastructure can lead to downtime.
- **Maintenance:** Requires complete shutdown for maintenance or repairs.

2.2.2 Tier II: Redundant capacity components

Tier II data centers provide better reliability compared to Tier I by incorporating some redundant components, such as backup power and cooling systems. However, they still rely on a single distribution path. They are ideal for small to medium-sized businesses with moderate uptime requirements. The main characteristics of Tier II data centers are described below:

- **Uptime:** 99.741% (22 hours of annual downtime).
- **Redundancy:** Power and cooling systems that have some redundancy.
- **Fault Tolerance:** Limited fault tolerance; failures in non-redundant components can still cause downtime.
- **Maintenance:** May require partial shutdowns for maintenance.

2.2.3 Tier III: Concurrently maintainable

Designed to support more critical operations, offering a higher level of redundancy and reliability. They feature multiple independent distribution paths for power and cooling, allowing maintenance or repairs without disrupting operations. They are suitable for enterprises and organizations with mission-critical applications that require high availability.

- **Uptime:** 99.982% (1.6 hours of downtime per year).
- **Redundancy:** N+1 redundancy, that is, there is at least one backup component for every critical system.
- **Fault Tolerance:** Can sustain at least one failure or maintenance activity without causing downtime.
- **Maintenance:** No shutdown is required for maintenance or repairs.

2.2.4 Tier IV: Fault-tolerant

Tier IV data centers are the most advanced and reliable, designed for organizations with zero tolerance for downtime. They feature fully redundant systems and multiple independent distribution paths, ensuring continuous operation even in the event of a failure. They are ideal for large companies, financial institutions, and government agencies that require maximum uptime and reliability.

- **Uptime:** 99.995% (26.3 minutes of downtime per year).
- **Redundancy:** 2N+1 redundancy,which means that there are at least two independent backup systems for every critical component.
- **Fault Tolerance:** Fully fault-tolerant; can sustain multiple failures without causing downtime.
- **Maintenance:** No shutdown is required for maintenance or repairs.

2.3 Key Components and Energy Management in Data Centers

A data center can vary significantly in size, ranging from a single room to an entire building. However, regardless of their scale, the essential components remain consistent across different configurations. At the core of every data center are two primary categories of components: IT infrastructure and supporting infrastructure. The IT infrastructure includes servers, storage systems, and networking equipment, which are responsible for processing, storing, and transmitting data. On the other hand, the supporting infrastructure consists of several critical systems that ensure the smooth operation of the data center. These include electrical power systems, such as UPS and backup generators, and other; mechanical or cooling systems, such as air conditioning units, to maintain optimal temperature and humidity levels; physical security measures, such as access controls and surveillance cameras, to protect sensitive equipment and data; and cabling and racks, which efficiently organize and connect IT equipment. Together, these components work seamlessly to ensure the reliable and efficient operation of the data center [1].

2.3.1 IT infrastructure

IT Equipment systems are the backbone of any data center, which includes servers, storage systems, and networking devices, collectively called *critical loads*. The load of IT equipment is dynamic and fluctuates throughout the day depending on the activity of the users. For example, during peak usage times, such as business hours or significant events, the demand for processing power and storage increases significantly. In contrast, during off-peak hours, the load decreases, which requires effective management strategies to enhance performance and energy consumption [35]. The key energy-consuming components of data centers are described in the following.

- **Servers:** Servers are the backbone of data centers, playing a crucial role in processing data and delivering outputs. These machines consume significant amounts of energy, generating substantial heat in the process, which requires the implementation of effective cooling systems [41]. The energy consumption of the servers

is mainly influenced by several components, including the CPU, memory, cooling fans, and input/output (I/O) devices. Among these, the CPU is typically the most power-intensive component, as its performance is directly correlated with energy consumption [42]. As server workloads increase, particularly with the increase of multi-core processors, overall energy consumption escalates significantly, leading to higher operational costs and higher demands on cooling infrastructure [36].

Research indicates that optimizing server efficiency and workload management can lead to substantial reductions in energy consumption, highlighting the importance of energy-sensitive design and operational practices in data center environments [42]. Furthermore, advances in server technology, such as energy-efficient processors and improved power management techniques, are essential to mitigate the environmental impact of data centers while maintaining performance and reliability.

- **Storage:** Data centers employ primary and secondary storage solutions to efficiently manage their data. Primary storage is concerned with short-term processing or access, and performs at the speed needed for applications or workloads that need to respond quickly. In contrast, secondary storage, which encompasses hard disk drives (HDDs) and solid-state drives (SSDs), is employed for long-term data retention. HDDs operate using mechanical components that read and write data on spinning disks, resulting in slower access times compared to their solid-state counterparts. Conversely, SSDs make use of flash memory technology, allowing significantly faster data access and improved performance, particularly for high-demand applications [11].

  Storage systems are considerable power consumers, ranking second only to servers in energy usage within data centers. Research indicates that as data storage needs grow, so does the energy consumption associated with these systems. For example, SSDs generally consume less power than HDDs, especially during read and write operations, making them a more energy-efficient option for many data center applications [44]. However, the overall energy impact of storage solutions remains substantial, requiring effective management strategies to improve energy efficiency in the data center environment [41].

- **Networking:** Networking devices are vital to facilitate communication between various components with a data center, as well as to connect the data center to external networks. These devices include switches, routers, firewalls, and load balancers. Switches connect multiple devices within the same network segment, allowing them to communicate with each other efficiently by forwarding packets of information based on MAC addresses. This capability is essential for maintaining high-speed connectivity among servers and storage systems. Routers, on the other hand, connect different networks together and direct traffic between them based on IP addresses. They manage incoming and outgoing traffic by determining the optimal route for data packets to take through interconnected networks [45]. In accordance with preset security standards, firewalls monitor all incoming and outgoing communication to offer security; they help protect sensitive information from unauthorized access or cyber threats.

  In addition, load balancers distribute incoming network traffic across multiple servers to prevent any one server from receiving too many requests. This not only improves performance, but also improves reliability by providing redundancy; If one server fails, traffic may be routed to other servers that are up and running without disrupting service. Networking devices also affect energy consumption within a data center as they require power to operate continuously. Efficient network design can help minimize energy consumption by optimizing traffic flow and reducing the number of active devices needed at any given time. Implementing energy-efficient networking technologies such as low-power switches or routers can further enhance overall energy management strategies [46].

2.3.2 Supporting Infrastructure

The support infrastructure of data centers comprises two main parts: electrical components, which ensure reliable power supply through systems such as UPS and rectifiers, and mechanical components, which manage temperature and airflow using cooling systems and CRAC units. Together, these components maintain optimal operational conditions and protect sensitive IT equipment.

A. Electrical Power Systems: Are critical for the continuous operation of data centers, ensuring a reliable and uninterrupted power supply to support essential IT functions. These systems typically encompass a variety of components, including diesel generators, transformers, power distribution frames, rectifiers, batteries, lighting systems, and uninterruptible power supply (UPS) systems. For example, UPS systems and backup generators are vital to guarantee power availability during power outages by providing immediate power to IT equipment. This capability minimizes downtime and protects sensitive equipment from interruptions, allowing operations to continue smoothly even in the event of a power failure. The design of these electrical power systems handles the fluctuating loads produced by IT equipment throughout the day and ensures an effective energy distribution [16][32].

- **Uninterruptible Power Supply (UPS) Systems**

  UPS systems are crucial for modern infrastructure, particularly in data centers and facilities that require continuous power supply. They act as a bridge between utility power and IT equipment, providing backup power to ensure uninterrupted operation during outages or disturbances. This capability minimizes downtime and protects against data loss and hardware damage, making UPS systems crucial to maintaining operational continuity.

  The main purpose of a UPS is to serve as a backup power source. In case of power failure, the UPS immediately supplies power, allowing the connected equipment to continue to operate consistently. In addition, UPS systems condition power by filtering and stabilizing the incoming electrical supply, correcting voltage fluctuations, surges, and sags to ensure that sensitive electronics receive clean and reliable power. Energy storage is another vital aspect, with UPS systems storing energy in batteries for immediate use when mains power is interrupted.

  The power conversion process within a UPS involves several key steps. Incoming alternating current (AC) power from the utility is first converted to direct current (DC) by the rectifier, which charges the batteries and provides a stable voltage for the inverter. When mains power fails, the inverter converts the stored DC power back to AC, supplying the connected IT equipment. This double conver-

sion process ensures that the output power remains consistent and reliable, which is critical for sensitive applications.

Design considerations for redundancy in UPS configurations are vital to maximizing availability and reliability. Common approaches include N+1, configurations that add one additional UPS unit to provide backup, and 2N configurations that involve having duplicate systems for uninterrupted power supply [47]. Proper load distribution across multiple UPS units is essential to prevent overload and enhance overall performance. Many modern UPS systems also feature hot-swap capability, allowing maintenance or replacement of units without shutting down the entire system. In addition, monitoring and management features provide real-time data on battery performance and health, facilitating proactive maintenance.

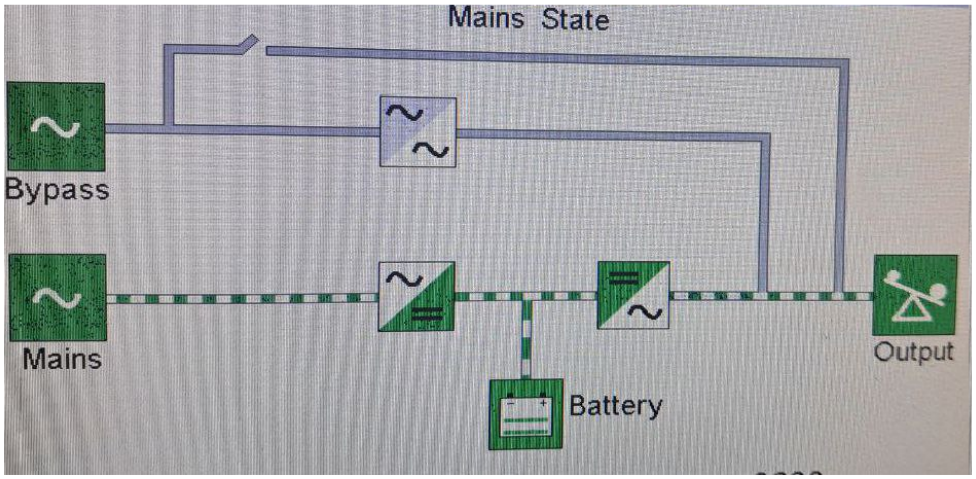

Figure 3: Nefas silk double conversion UPS system design

As shown in Figure 3 the double conversion UPS system is designed to provide high availability and reliability for connected loads by continuously converting the incoming power and ensuring clean and stable output. The primary components of a double conversion UPS system include the rectifier, inverter, static bypass switch, manual bypass switch, and battery, each playing a critical role in maintaining power continuity in the event of a failure or outage [11].

The rectifier serves as the first line of defense by converting the incoming AC from the utility power into DC. When utility power is available, the rectifier takes the AC input and transforms it into DC power, which is used to charge the batteries and supply the inverter for conversion back to AC. This component also regulates

voltage and current, ensuring that batteries are charged appropriately and that the system operates efficiently. The inverter is another crucial component, responsible for converting the stored DC power from the batteries back to AC power for the connected loads. In the event of power failure or disturbance, the inverter is taken over to provide power to the load. Continuously delivers clean and stable AC power that meets the requirements specifications for sensitive equipment. This component ensures that there is no transfer time during the switch from mains power to battery, providing seamless power continuity. The static bypass switch allows for the instantaneous transfer of the load back to the utility supply in the event of an inverter failure or overload. This switch monitors the performance of the inverter and can seamlessly switch the load to the mains supply without interrupting the power to the connected equipment. In the event of an inverter fault or if the UPS requires service, the static bypass switch ensures that the load continues to receive power directly from the mains, thus maintaining uptime.

In addition, the manual bypass switch provides technicians with the means to divert the load directly to the power supply for maintenance purposes without interrupting the power supply to the connected equipment. This switch is essential during routine maintenance or repairs, allowing safe servicing of the UPS while ensuring that connected loads remain powered. It gives technicians a physical means to isolate the UPS from the load, facilitating safe maintenance procedures [48].

Lastly, the battery stores energy to provide backup power during outages or when the mains supply is unstable [49]. The batteries are charged by the rectifier when the utility power is available,and the power is discharged through the inverter when needed. They serve as the main energy source during power failure, ensuring that there is no interruption in the power supply to critical loads. In addition, the battery's condition and functionality are tracked by the battery management system , ensuring optimal operation and longevity. Together, these components work in harmony to ensure high availability and reliability for connected loads. The continuous operation of the rectifier charges the batteries while supplying power to the inverter,providing a stable AC output at all times. When a power disturbance occurs, the inverter immediately takes over, supplying power from the batteries without any transfer time, which is crucial for maintaining up-time for sensitive

equipment. The static bypass switch adds an extra layer of protection by allowing the load to be powered directly from the mains in case of inverter failure, while the manual bypass switch enables safe maintenance of the UPS without service interruptions.

UPS Battery Configuration:

UPS systems utilize a series-parallel arrangement of batteries to achieve the required voltage and capacity for various applications. In this configuration, batteries are connected in series to increase voltage and in parallel to increase capacity. The number of connected batteries depends on factors such as the voltage of the individual battery cell, the desired system voltage, the actual loads, and the expected power duration. For example, a 12V system typically requires six 2V cells in series, while parallel connections are added to meet the capacity needs based on load requirements and desired backup time [**?**].

Common types of UPS batteries include valve-regulated lead acid batteries (VRLA), sealed cells and flooded cells, each with distinct advantages and disadvantages in terms of maintenance, cost, and useful life. The 12V 200Ah battery is a popular choice for UPS systems, providing a nominal voltage of 12 volts and the ability to deliver 20 amps for 10 hours. This significant capacity makes it suitable for backup power applications, ensuring critical equipment remains operational during outages. Understanding these configurations and battery types is essential for optimizing UPS performance and ensuring reliable energy storage for critical loads.

UPS Operation Working Modes:

- **Normal Mode (Mains State):** This mode is active when the UPS is receiving stable power from the utility grid. The system operates normally, providing power to the load while simultaneously charging the batteries [11].
- **Bypass Mode:** The display shows a "Bypass" indicator, which suggests that the UPS can operate in bypass mode. In this mode, power flows directly from the mains to the load, bypassing the UPS. This is usually done for maintenance or when the system is experiencing problems [49].

– **Battery Mode:** If there is a power failure or significant disturbance, the UPS switches to battery mode, drawing power from its internal batteries to continue supplying the load. The monitor would indicate the battery status during this mode [48].

- **Rectifier**

Rectifiers are essential in data centers, as they convert alternating current (AC) to direct current (DC) for critical components such as servers, networking equipment, and storage systems. By providing stable power, they ensure optimal performance and compatibility while also reducing operational costs and energy consumption. High-efficiency rectifiers minimize energy losses during the AC-to-DC conversion process, making them particularly important for continuous operations required in a 24/7 environment.

In addition to providing a stable and regulated DC output that prevents fluctuations that could disrupt sensitive equipment, rectifiers often feature modular designs, such as those found in models such as the TP48, which allow for scalability [50]. This enables data centers to expand their power capacity as demand grows without significant disruptions. Advanced rectifiers also offer real-time monitoring capabilities for proactive maintenance, while built-in protection features protect the rectifier and connected equipment from potential electrical issues, ensuring operational continuity and reliability.

Rectifier Load Battery:

The 2V 3000Ah stationary valve-regulated lead acid battery (VRLA) is a crucial energy storage solution for data centers, supporting rectifier equipment [50]. With a nominal voltage of 2 volts and a capacity of 3000 ampere hours over a 10-hour discharge period, VRLA batteries ensure consistent power, ensuring servers and networking equipment remain operational during outages. The sealed construction with built-in valves minimizes maintenance and prevents electrolyte spillage, making VRLA batteries ideal for continuous operation in data centers.

B. Mechanical (Cooling) Systems: Cooling systems, also known as mechanical infrastructure, are essential components of data centers, designed to maintain optimal operating temperatures for IT equipment and ensure efficient performance. ethio telecom, like many other organizations, employs air-cooled systems in its data centers, which are critical to managing the significant heat generated by servers and other devices during operation. Effective cooling solutions are vital to prevent overheating, which can lead to equipment failure and reduced service life.

In air-cooled or legacy data centers, server racks are arranged in cold and hot aisles. In cold aisles, the front sides of the server racks face each other, allowing cool air to flow toward the servers. In contrast, the hot aisles are lined with the rear sides of the racks, where hot exhaust air exits. The chilled air produced by Computer Room Air Conditioning (CRAC) units is driven into the cold aisles, either through the floor plenum and perforated tiles in a raised floor design or through diffusers in the ceiling in non-raised floor setups. This design allows for effective air circulation by using the space beneath the floor for cabling and cooling distribution [37][32].

The arrangement of the hot and cold aisles is another vital aspect of the cooling system design. By positioning server racks in alternating rows, with cold air intakes facing one aisle and hot air exhausts facing the other, data centers can prevent air recirculation and optimize cooling performance. This configuration helps to ensure that cool air is efficiently directed to the servers, while hot air is expelled away from them. According to the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), maintaining a temperature between 18 and 27 degrees Celsius with a relative humidity of 20 to 60 percent is crucial for operational efficiency [52].

Recent advances in cooling technologies have focused on improving energy efficiency

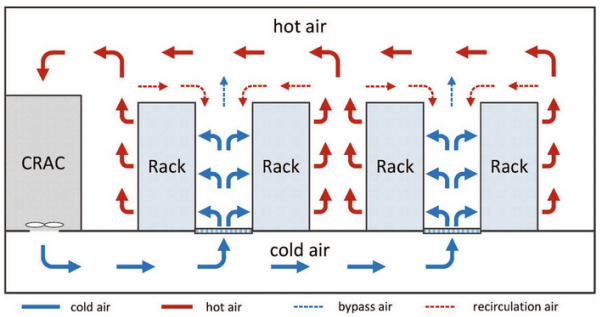

Figure 4: The cold/hot aisle design of a raised-floor data center [5]

while maintaining effective temperature control. For example, in-row cooling systems place cooling units directly between server racks, providing targeted cooling where it is needed most. This approach reduces the distance that cool air must travel, improving efficiency and reducing energy costs. In addition, liquid cooling systems have gained popularity due to their ability to absorb heat more effectively than traditional air-based systems. These systems use liquid coolant to remove heat from high performance components, such as CPU and GPU, allowing data centers to support more powerful hardware without compromising thermal management [54].

In addition, innovative control strategies have been developed to optimize the operation of cooling systems. Artificial intelligence (AI) and machine learning algorithms can analyze data from real-time sensors throughout the data center to dynamically adjust cooling output according to current load conditions [37]. This adaptive approach not only enhances cooling efficiency but also contributes to significant energy savings by reducing unnecessary cooling when demand is low.

2.4 Design Considerations for Data Center Power Supply Systems

The power supply path in the data center of the Nefas Silk core site designing an efficient and reliable data center power system requires careful planning from the initial power input to the final distribution to IT equipment. The process begins with the primary power source, typically a high-capacity 1350KVA main supply, which provides three-phase AC power to the facility. This power first passes through an ATS, ensuring a seamless transition to backup generators or alternate sources during grid failures, thus maintaining uptime. From the ATS, power flows to the Main AC Power Distribution Board (PDB) (600A/3P), which acts as the central hub to distribute electricity throughout the data center.

The power system then divides into two critical paths: AC-based distribution for supporting infrastructure and DC-based distribution for IT loads. On the AC side, the power is routed to the UPS (240KVA) via a UPS PDF, ensuring uninterrupted power

to sensitive IT equipment during short-term outages. The UPS output feeds AC loads, such as servers and networking equipment, while also connecting to a battery bank for energy storage. Simultaneously, another branch powers the Air Conditioner PDF, which supports the cooling systems essential to keep the ideal operating temperatures. On the DC side, the Rectifier converts AC power from the Main AC PDF to 3000A DC, which is then distributed through the DC PDF to power DC-based IT equipment. The rectifier also charges the battery system, providing backup power during extended outages. This dual-path design (AC and DC) enhances redundancy and efficiency, ensuring continuous operation even if one system fails.

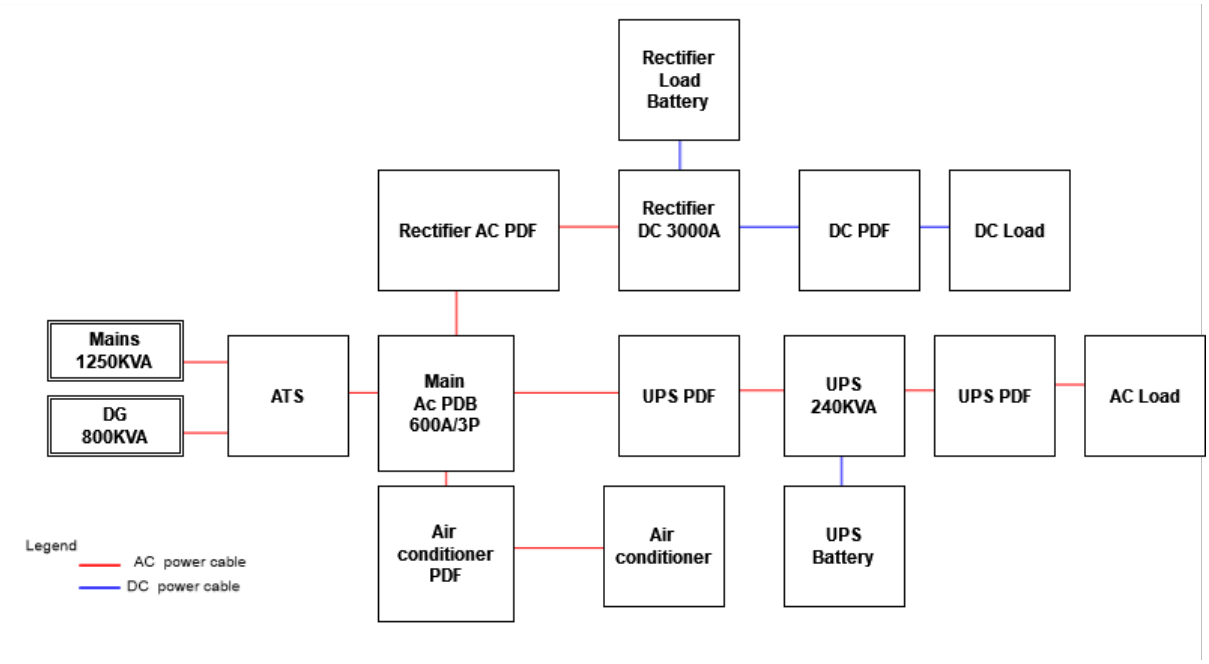

Figure 5: Power delivery path of a Nefas silk data center layout design.

2.5 Evaluating Energy Efficiency Metrics and Performance

Energy efficiency performance in data centers is evaluated using various metrics that assess how efficiently energy is used for computing tasks versus supporting infrastructure. Key metrics include PUE, which measures the total energy consumed by the facility relative to the energy used by IT equipment, and Data Center Infrastructure Efficiency (DCIE), which expresses the proportion of energy used for useful work compared to total energy consumption. These metrics provide insights into operational efficiency, guiding data center operators in identifying inefficiencies and implementing best practices for

energy management. By continuously monitoring and optimizing these metrics, data centers can reduce operational costs, enhance sustainability, and improve overall performance in an increasingly energy-conscious landscape.

Efficiency Metrics in Data Centers

1. **Power Usage Effectiveness (PUE)**: Is a measure of the efficiency with which a data center uses energy. Calculated by dividing the total energy consumed by the data center (including all supporting systems) by the energy used solely for IT equipment. A lower PUE shows better energy efficiency, with ideal values that typically range from 1.0 to 1.8; values below 1.2 indicate highly efficient data centers [34].

   This analysis uses a comprehensive assumptions to calculate PUE for a Nefas Silk data center design, focusing on total facility power and IT equipment power, accurately reflecting the facility's power distribution architecture as shown in Figure 5. The total facility power is measured at the critical input point of the Main AC PDB, which consolidates all three-phase power (Phases A, B, and C) after it passes through the ATS. The whole facility's energy footprint is captured by this assessment, which includes IT systems as well as support infrastructure like lighting, cooling, and the inevitable power conversion losses in UPS systems and rectifiers. For the IT equipment power, the calculation specifically tracks the energy delivered to computing resources through two primary pathways. This component is defined as the sum of the output power from rectifiers and the UPS. More specifically, it includes power for AC-powered devices via the UPS output and DC-powered equipment through the rectifier outputs.

   This careful distinction between active computational power and support infrastructure consumption is fundamental for obtaining an accurate PUE measurement. The methodology acknowledges that while all incoming power contributes to total facility consumption, only the portion directly energizing IT equipment represents useful computational work. This approach ensures a precise assessment of energy efficiency within the data center.

Calculation of PUE

$$\text{PUE} = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}} \tag{2.1}$$

Where:

$$\text{Total Facility Power} = \sum_{j=1}^{3} \text{Active power phase } j \quad (\text{where } j = A, B, C)$$

$$\text{IT Equipment Power} = \text{Rectifier output power} + \text{UPS output power}$$

The formula for calculating PUE can also be expressed as:

$$\text{PUE} = \frac{\sum_{j=1}^{3} \text{Active power phase } j}{\text{Rectifier output power} + \text{UPS output power}} \tag{2.2}$$

2. **Data Center Infrastructure Efficiency (DCIE)**: Measures the efficiency of a data center infrastructure to use energy for computing tasks. It is calculated as the percentage of energy used for useful work compared to total energy consumption. A higher percentage of DCIE indicates a more efficient use of energy for computing, with values above 80% generally considered excellent, while lower values highlight potential inefficiencies in the infrastructure[34].

Calculation of DCIE

$$\text{DCIE} = \left(\frac{\text{Energy Used for Useful Work}}{\text{Total Energy Used by Data Center}} \right) \times 100 \tag{2.3}$$

These metrics are essential for data center operators seeking to optimize energy use, reduce operational costs, and improve overall sustainability.

Table 1: PUE and DCIE values and the level of efficiency source [10]

PUE	DCIE	Level of Efficiency
3.0	33%	Very Inefficient
2.5	40%	Inefficient
2.0	50%	Average
1.5	67%	Efficient
1.3	83%	Very Efficient

2.6 Root Cause Analysis in Data Center Operations

Root Cause Analysis (RCA) is a systematic approach that is used to identify the underlying causes of problems or inefficiencies within data centers. By focusing on root causes rather than symptoms, organizations can implement effective solutions that prevent recurrence and enhance overall operational efficiency. In the context of data centers, RCA can be applied to various areas, including energy consumption, system failures, performance issues, and operational inefficiencies. Key terms related to data center concerns are provided here.

Energy Inefficiency:

Energy inefficiency in data centers is often indicated by high power consumption efficiency (PUE) values, excessive energy bills, and increased carbon footprint. The potential root causes of this inefficiency include inefficient cooling systems that do not align with the actual heat load generated by the IT equipment, resulting in inadequate cooling or excessive energy consumption. Furthermore, poor airflow management can lead to hot spots within the facility, which requires overcooling in certain areas while other parts remain under-cooled[13]. Furthermore, overprovisioning of IT resources where more servers and equipment are deployed than necessary can lead to low utilization rates, causing a significant waste of energy,as these underutilized resources still consume power without actually contributing to operational needs[57]. Addressing these issues is crucial for improving energy efficiency and reducing operational costs in data centers.

System Failures:

Data center system failures are critical issues that can cause frequent hardware failures, unexpected downtime, and data loss incidents. These failures often arise from

several potential root causes[59]. One significant cause is the lack of regular maintenance schedules for critical infrastructure, which can result in undetected wear and tear on equipment, leading to breakdowns. Furthermore, inadequate redundancy measures for power and cooling systems can leave data centers vulnerable to outages; if a primary system fails without a reliable backup, it can lead to significant operational failures [39]. Aging equipment that has surpassed its operational lifespan is another contributing factor, as older systems may not perform reliably under current demands and may be more prone to failure. Addressing these root causes through proactive maintenance, improved redundancy planning, and timely equipment upgrades is essential to improve the reliability and resilience of data center operations [41].

Performance Issues:

Data centers face performance issues such as slow response times to applications, increased latency, and user complaints about service quality, significantly impacting user experience and operational efficiency. These problems can arise from several potential root causes. One major factor is network bottlenecks, which occur when there is insufficient bandwidth or when outdated network equipment is unable to handle current traffic demands [60]. Additionally, inefficient resource allocation among virtual machines or containers can lead to some workloads being starved of necessary resources, resulting in degraded performance for critical applications. Furthermore, configuration errors in software or hardware settings can alter optimal functioning, causing delays and inconsistencies in service delivery [59]. Addressing these root causes through network upgrades, improved resource management practices, and meticulous configuration audits is essential to enhance the overall performance of data center operations.

Operational Inefficiencies:

Data centers often experience operational inefficiencies, characterized by high operational costs, low staff productivity, and increased time spent troubleshooting problems. These inefficiencies often arise from several key root causes. A significant factor is the lack of automation in the monitoring and management processes, which can lead to manual errors and slow response times when addressing operational issues [59]. Further-

more,insufficient training for staff on best practices in data center operations can result in ineffective use of resources and suboptimal performance, as employees may not be fully equipped to handle the complexities of modern data center environments. Furthermore, poor documentation of processes and procedures can create confusion among team members, leading to inconsistent practices and increased reliance on trial-and-error methods. By addressing these root causes through the implementation of automation tools, comprehensive training programs, and improved documentation practices, data centers can improve operational efficiency, reduce costs, and improve overall service delivery.

Importance of RCA in Data Centers

RCA plays a crucial role in improving energy efficiency within data centers by systematically identifying and addressing the underlying causes of energy-related problems [61]. One of the primary benefits of RCA is its ability to pinpoint inefficiencies in energy consumption, such as excessive power use due to inadequate cooling systems or overprovisioned IT resources. By focusing on these root causes, data center operators can implement targeted solutions that not only reduce energy waste but also optimize resource utilization, leading to significant cost savings [63][64][25].

In addition, RCA supports proactive decision-making with respect to energy management. By analyzing historical data and identifying patterns related to energy consumption, organizations can forecast potential problems before they escalate. This predictive capability allows for timely interventions, such as adjusting cooling loads or redistributing workloads across servers, which helps maintain optimal operating conditions while minimizing energy use [11].

RCA also fosters a culture of continuous improvement in energy practices. By encouraging teams to learn from past incidents and adapt their processes accordingly, data centers can improve their operational efficiency over time [10]. This iterative approach not only improves current energy management strategies, but also promotes collaboration between different teams, as effective RCA often requires input from various stakeholders, including facilities management and IT operations. In addition, implementing RCA can lead to better compliance with regulatory standards related to energy efficiency. As data centers face increasing scrutiny regarding their environmental impact, having a transparent RCA process helps organizations demonstrate their commitment to sustainability and responsible energy management.

3 Deep Learning and SHAP

This chapter presents the ideas behind Deep Learning (DL) and the SHAP frame-work which are essential to understand their role in data analysis and model interpretation. The chapter also deals with the advancement of these methods to determine where the source of energy inefficiency of the data centers lies using CNNs and SHAP for interpretability.

3.1 Core Concepts of Machine Learning

Machine learning is a part of artificial intelligence (AI) that is basically concerned with the creation of algorithms that have the potential to learn from data and then make predictions based on those data. It encompasses various tasks, including speech recognition, decision-making, and problem-solving. Within this field, ML can be categorized into three primary types of learning: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning utilizes labeled data to train models, enabling them to learn function approximations under the guidance of a "teacher." This approach is particularly effective for classification and regression tasks. In contrast, unsupervised learning deals with unlabeled data, allowing exploratory data analysis to uncover hidden patterns or groupings. Reinforcement learning involves agents that learn optimal behaviors through interactions with their environment, guided by a predefined reward function, and is commonly used in scenarios requiring sequential decision-making.

To identify the most suitable model for a given problem, it is essential to evaluate and compare multiple models. This process involves evaluating their performance based on relevant metrics and ensuring that the chosen model aligns with the specific requirements of the task at hand. Using machine learning in conjunction with root cause analysis significantly enhances the ability to identify and address inefficiencies. By applying supervised and unsupervised learning techniques, organizations can analyze operational data to detect anomalies and predict potential failures. This data-driven approach improves the accuracy of RCA and facilitates proactive decision-making, ultimately leading

to enhanced operational efficiency and sustainability in various sectors, including manufacturing and data management.

3.1.1 Deep Learning Models

DL, or Deep Neural Networks (DNN), is a specialized subset of ML that uses ANN to analyze complex data. It is inspired by the human brain that allow machines to identify patterns and make predictions with stunning accuracy [65][58]. These provide many advantages over traditional approaches, one of the main advantages of these approaches, is their capability to automatically learn features from the raw data, thus being less dependent on manual feature engineering. This capability enables us to capture complex,unstructured, non-linear relationships within the data, making them particularly effective for datasets where patterns do not follow straightforward trends.

In addition,deep learning models maintain the memory of previous time steps, effectively using past observations to make more accurate future predictions. They also demonstrate robustness to noise and outliers by employing techniques like dropout and regularization, enhancing their generalization to unseen data. Overall, these advantages position deep learning as a transformation tool for time series analysis, driving improvements in accuracy and insights across various applications.

Key Characteristics of Deep Learning

1. Hierarchical Feature Learning: Deep learning models, especially DNN are designed to represent data in several layers of abstraction. Each layer is a different level of features so the model is able to learn and represent complex data patterns.Lower layers usually find simple features (e.g., edges in images), and higher layers use these to create new complex representations (e.g., shapes or objects). Such a hierarchical learning process is the reason why deep learning models can efficiently manage high- dimensional and unstructured data.

2. Layered Architecture: ANN are the core of deep learning, providing the framework for data processing and learning. They consist of interconnected layers of nodes (neurons)

that work together to analyze data. The neural network architecture, that is how many layers and how many neurons per layer, is an important factor in their performance, capacity , and ability to learn complex patterns from data [55].
The basic structure of an ANN, as illustrated in Figure 1, is organized into three main layers.

- **Input Layer**: This is the basic layer where the data set is fed into the model. The input layer neurons are in a one to one relation with the features of the data set. To illustrate, in a classification task, each feature of the input object is a neuron. The input layer does not carry out any operations; it just moves the information to the following layer.

- **Hidden Layers**: These in-between layers, which are intermediate between the input and output layers, utilize activation functions to change the information obtained from the previous layer. ANNs are capable of having one or several hidden layers, each made up of numerous neurons. The structure, which includes depth (amount of hidden layers) and width (neurons per layer), defines the model's ability to learn complex patterns and generalize them to new data.

- **Output Layer**: The last layer generates the output based on the features that have been learned from the intermediate layers. For a classification problem, the number of neurons in the output layer corresponds to the number of categories.

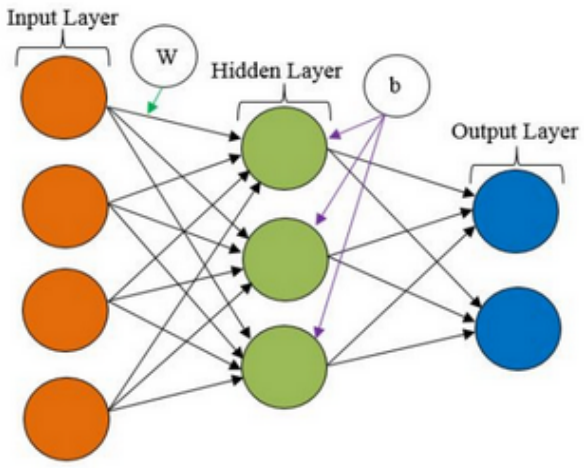

Figure 6: Basic structure of an artificial neural network [6]

Weights and Biases: In an ANN, weights (W) and the biases (b) are critical components that determine how inputs are transformed as they pass through the network layers. The connections between neurons are given numerical values called weights, which affect the signal's strength and direction. A bias is an extra parameter that is applied before the activation function is applied to the weighted sum of inputs. This allows the model to adjust the output independently of the input values, improving its ability to learn complex patterns and relationships within the data.

Mathematical Representation

The output of a neuron can be mathematically represented as:

Weighted Sum of Inputs:

$$z = \sum_{i=1}^{n} W_i \cdot x_i + b \tag{3.1}$$

Where:

- z is the weighted sum.
- W_i represents the weight associated with the i-th input.
- x_i is the i-th input value.
- b is the bias term.

3. Automatic Feature Extraction: DL models automatically extract relevant features from raw data, unlike traditional ML approaches that often require manual feature engineering. This automation enables DL to identify complex, non-linear relationships within the data, enhancing model performance, and reducing the time and effort needed for feature selection.

4. Nonlinearity through activation functions: In order to provide nonlinearity to the model and enable it to recognize and depict intricate patterns, neurons employ activation functions [65]. The output a of the neuron is then calculated by applying an activation function f:

$$a = f(z) = f\left(\sum_{i=1}^{n} W_i \cdot x_i + b\right) \tag{3.2}$$

The selection of activation function has a major impact on the way networks train on the data and their overall performance. Common activation functions include

- **Sigmoid Function**: This function restricts the output values between 0 and 1, making it useful for binary classification tasks. It is defined as:

$$O = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.3}$$

- **Rectified Linear Unit (ReLU)**: This function outputs zero for negative inputs and returns the input itself for positive values. ReLU is widely used due to its simplicity and effectiveness in mitigating problems such as vanishing gradients. It is defined as:

$$O = \text{ReLU}(z) = \max(0, z) \tag{3.4}$$

- **Hyperbolic Tangent (tanh)**: This function keeps output values within the range of [-1, 1], providing a zero-centered output that can lead to faster convergence during training. It is defined as:

$$O = \tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{3.5}$$

5. Backpropagation and Optimization: DL models use backpropagation to update weights according to the prediction error. By adjusting the weights of neurons to minimize the loss function,which measures the difference between the expected and actual output, this procedure is essential to neural network training. Popular optimization algorithms include:

- **Stochastic Gradient Descent (SGD)**: A common optimizer that uses a subset of training data to update weights.
- **Adaptive Moment Estimation (Adam)**: preferred in deep learning because of its flexible learning rates, which improve optimization and speed up training convergence.
- **AdaGrad**: An optimizer that adjusts learning rates based on past gradients for each parameter.

The choice of optimizer can significantly impact how effectively a neural network learns from data and how quickly it converges to an optimal solution.

6. Hyperparameters: Finding the optimal hyperparameters for a DL model requires experimentation and validation.Key hyperparameters commonly used in deep learning include

- **Learning Rate**: Controls the step size during optimization. A high rate may lead to suboptimal convergence, while a low rate can slow down training.
- **Number of Hidden Layers**: Affects the depth of the network. More layers can capture complex patterns but may increase the risk of overfitting.
- **Number of Neurons per Layer**: Defines the network width. More neurons enhance learning capacity, but can lead to overfitting with limited data.
- **Number of Epochs**: Indicates complete passes through the training data. If the number of epochs is too low, the result may be an underfitting model, whereas if too many epochs are used, overfitting may occur.
- **Batch Size**: Finds how many samples are processed before updating the model parameters. Smaller sizes improve gradient estimates but increase training time, whereas larger sizes speed up training but may reduce accuracy.
- **Dropout Rate**: A regularization technique that randomly deactivates neurons during training to prevent overfitting. A higher rate improves generalization,but can hinder learning if too high.

Proper tuning of these hyperparameters is essential for effective convergence and generalization. Techniques like grid search and random search are commonly used for optimization.

3.1.2 Major Categories of Deep Learning Architectures

Deep learning architectures can be broadly classified into several main categories, each is appropriate for particular kinds of applications, data structures, and activities. Primary categories include recurrent neural networks (RNN), long- and short-term memory networks (LSTM), and convolutional neural networks (CNN). Below is a detailed look at each category [22]:

- **RNN**: designed to handle sequential data, effectively capturing temporal dependencies by maintaining a memory of previous inputs. They are particularly useful for tasks such as predicting future values based on past observations. RNNs may keep some form of memory because of their loops, which let data carry over from one step to the next. This characteristic makes RNNs suitable for tasks involving time-series data or natural language, as they can consider previous input when generating outputs. However, RNNs can struggle with long-term dependencies due to the vanishing gradient problem, which can hinder their performance on longer time series sequences [58].

- **LSTM**: is a specialized type of RNN designed to address the problem of vanishing gradients, allowing them to learn long-term dependencies effectively. They are widely used in complex time series forecasting and sequential prediction tasks, such as stock price forecasting and weather prediction. Although LSTMs are particularly effective for tasks that require context retention in longer sequences,such as language translation and sentiment analysis,they are computationally intensive and require more extensive training data and time, making them less efficient for simpler tasks[58].

- **CNN**: is a specialized architecture designed to process grid-like data, particularly object detection and image recognition. They utilize convolutional layers that apply filters (kernels) to the input data to automatically detect patterns, such as edges and textures, while reducing dimensionality through pooling layers. The hierarchical structure of CNN allows them to learn increasingly complex features at each layer, making them highly effective for visual recognition tasks [55][58].

3.2 Architectures of Convolutional Neural Networks (CNN)

CNN are a class of DL architectures that are particularly effective in analyzing visual and sequential data. They can automatically and adaptively learn spatial feature hierarchies from input data, which makes them ideal for applications like image recognition, object detection, and time series analysis. CNN consists of several essential building blocks that work together to process data and make predictions. These components include the input layer, convolutional layers, pooling layers, fully connected layers, and the output layer [7].

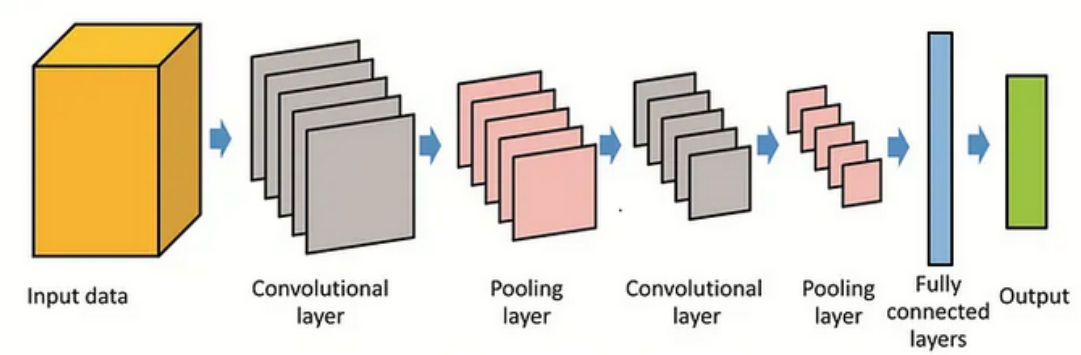

Figure 7: The architecture of a simple CNN model[7]

3.2.1 Component Layers in CNN

1. Input layer : The input layer receives raw data, which can be structured as a 1D array (for sequential data such as time series) or a 2D matrix (for image data). For time series data, the input is typically represented as a sequence of time steps, where each time step contains one or more features. Proper formatting of the input data is crucial for effective processing and feature extraction.

2. Convolutional layers : The core component of CNN is the convolutional layers, which are in responsible for identifying relevant features in the input data. These layers apply filters (kernels) to the input, sliding them across the data to detect patterns such as edges, textures, or temporal trends. The convolution operation involves computing the dot product between the filter and segments of the input data, producing a feature map that highlights detected features. Mathematically, the convolution operation can be represented as:

$$(I * K)(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} I(i+m, j+n) K(m,n) \tag{3.6}$$

Where:

- $I(i,j)$ is the value at position (i,j) in the input data.
- $K(m,n)$ is the value of the kernel at position (m,n).
- k is the size of the kernel (for example, for a 3×3 kernel, $k = 1$).

Following convolution, activation functions such as ReLU are applied to introduce non-linearity, allowing the model to learn complex patterns. Convolutional layers also use shared weights in different spatial locations, reducing the number of parameters and improving computational efficiency [19].

3. Pooling layer : The pooling layers downsample the feature maps generated by the convolutional layers, reducing dimensionality and computational complexity while mitigating overfitting. The two most common types of pooling are max pooling and average pooling. For a given input feature map I and a pooling window defined by size $p \times p$, max pooling can be represented as:

$$P(i,j) = \max_{m,n} I(i+m, j+n) \tag{3.7}$$

In contrast, average pooling computes the average value within the same window:

$$P(i,j) = \frac{1}{p^2} \sum_{m=0}^{p-1} \sum_{n=0}^{p-1} I(i+m, j+n) \tag{3.8}$$

Pooling serves several important purposes.

- **Dimensionality Reduction**: Reduces the number of parameters and computations, leading to faster training and inference.

- **Feature Extraction**: Retains the most salient features while discarding less important information.
- **Translation Invariance**: Provides some degree of invariance to translations in the input data, making the model more robust to variations.

4. Fully connected layer: Fully connected layers, also known as dense layers, classify data into different categories based on patterns learned from previous layers. The output of the last pooling layer is flattened into a vector and passed through one or more fully connected layers, where each neuron connects to every neuron in the previous layer. The output of a fully connected layer can be represented as:

$$Y = f(WV + b) \tag{3.9}$$

Where:

- Y is the output vector (for example, class scores).
- b is the bias vector.
- $f(\cdot)$ is an activation function applied to each element of the output vector, often used for binary classification.

These fully connected layers combine information from previous layers and make final predictions, integrating all extracted features to enhance predictive performance, especially in time-series analysis.

5. Output layer : The output layer produces the final classification or regression output. In classification tasks, this layer typically uses a softmax activation function to obtain probabilities for each class, allowing the model to make informed decisions based on the extracted features. For binary classification, the output layer consists of a single neuron with a sigmoid activation function, producing a value between 0 and 1 that represents the probability of the positive class.

3.2.2 Types of CNN

can be categorized according to the dimensionality of the input data: One-dimensional (1D), Two-dimensional (2D), and Three-dimensional (3D). Each type is designed to extract features from the data according to dimensionality [53].

- 1D CNN: Used for processing sequential data, such as time series or audio signals. They apply filters along a single dimension to capture temporal patterns effectively, making them suitable for tasks such as RCA and financial forecasting.
- 2D CNN: Commonly used for image processing,filters across height and width are applied to extract spatial hierarchies.
- 3D CNN: Extend 2D CNN by adding a third dimension (depth or time), making them ideal for analyzing volumetric data or sequences of images, such as video analysis or medical imaging.

3.3 Machine Learning Interpretability and Explainability

Interpretability and explainability have emerged as key concerns in the quickly changing fields of AI and ML. As ML models increasingly influence decision making processes in various sectors, particularly in sensitive applications such as healthcare, finance, and data center operations, the need for transparency has become paramount. Interpretability refers to the degree to which a human can understand the cause of a decision made by a model, while explainability encompasses a broader range of methods and techniques aimed at elucidating how these models operate.

1. Importance of Interpretability and Explainability

The demand for interpretability and explainability arises from the need for accountability in automated systems. When models are deployed in sensitive areas, such as predicting energy inefficiencies in data centers or diagnosing medical conditions, understanding the rationale behind decisions becomes crucial. Interpretability allows practi-

tioners to diagnose model performance, identify potential biases, and ensure ethical use of algorithms. In addition, it fosters trust and acceptance among stakeholders, as they are better equipped to comprehend and evaluate the results produced by ML systems.

2. Categories of Interpretability Methods

Interpretability methods can be classified according to various attributes, including model-specific versus model-agnostic methods, intrinsic versus post hoc interpretability, and global versus local interpretability [21][20].

- **Model-Specific Methods:** These methods are tailored to specific classes of models, using their inherent characteristics to provide explanations. For example, decision trees and linear regression models are often easier to interpret due to their straightforward structures. In contrast, deep learning models are typically more complex and require specialized techniques for interpretation.
- **Model-Agnostic Methods:** These methods can be applied to any machine learning model, regardless of its architecture. Techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) fall into this category, allowing flexible interpretability across various algorithms.
- **Intrinsic Interpretability:** This approach focuses on the model's design, explaining its behavior directly from its input to its parameters. Models such as linear regression and decision trees are inherently interpretable, as the relationships between inputs and outputs are explicit.
- **Post-Hoc Interpretability:** This involves analyzing complex models after they have been trained to derive explanations for their predictions. Techniques such as feature importance scores and visualizations help elucidate how model predictions are formed.
- **Global Interpretability:** These methods aim to explain the entire logic of a model, providing insight into how different features impact all possible outcomes. This helps stakeholders understand the general behavior of the model.
- **Local Interpretability:** These methods focus on individual predictions, explain-

ing how specific input features influence a particular outcome. This granularity is crucial for users who need to understand the reasoning behind specific decisions made by the model.

3.3.1 SHAP (SHapley Additive exPlanations) Framework

SHAP is a sophisticated method designed to improve the interpretability of machine learning models by providing information on the contributions of individual features to the predictions of a model. This method takes advantage of principles from cooperative game theory, specifically the Shapley value, to fairly distribute the prediction among the input features. SHAP offers local and global interpretability as well, thus being a multipurpose tool for understanding the complex models.

- **Local Interpretability**: At the local level, SHAP values explain individual predictions by showing how much each feature contributes to the prediction for a specific instance. This is particularly useful for understanding the behavior of the model in specific cases [62][56].
- **Global Interpretability**: On a broader scale, SHAP facilitates global interpretability by aggregating SHAP values across multiple instances. This allows practitioners to identify which features are generally influential in the model's decision-making process.

The SHAP value for a feature j for a specific instance x is calculated using the formula:

$$\phi_j(f, x) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \left(f(S \cup \{j\}) - f(S)\right) \tag{3.10}$$

Where:

- $\phi_j(f, x)$ is the SHAP value for the feature j, for instance x.
- N is the set of all the features.

- S is a subset of features that does not include j.
- $f(S)$ is the prediction of the model using the features in the subset S.

Percentage Contribution

To compute the percentage contribution of each feature to the model's output (e.g., for inefficiency predictions like high PUE), use the following formula:

$$\text{Percentage Contribution} = \left(\frac{\text{Feature SHAP Value}}{\text{Total Impact}}\right) \times 100 \tag{3.11}$$

Where:

- Feature SHAP Value: The SHAP value assigned to a specific feature .
- Total Impact: The sum of all SHAP values for all features.

This combined approach facilitates a deeper understanding of model behavior, allowing practitioners to identify and quantify the contributions of different features effectively.

Methods to Calculate SHAP Values

Different methods exist within the SHAP framework to calculate SHAP values, customized for different types of models [43]:

- **Tree Explainer**: Specifically designed for tree based models, the tree explainer leverages the tree structure to perform SHAP value computations efficiently, thereby enabling faster calculations by a large margin. Models such as XGBoost, LightGBM, and random forests frequently use this method. For example, in a credit scoring model using XGBoost, the tree explainer can provide insight into which features (e.g., income, credit history) have the most significant influence on the predicted creditworthiness of an applicant.
- **Linear Explainer**: This method computes SHAP values from linear models, providing information on how each feature contributes to the prediction based on their weights. It is used with linear regression and logistic regression models. For

instance, in marketing campaign analysis, the Linear Explainer can show how different advertising channels (e.g., social networks, email) impact the likelihood of customer purchases.

- **DeepExplainer**: Deep SHAP is specifically designed for deep learning models. It extends the traditional SHAP approach by incorporating a baseline input that serves as a reference point to quantify the contribution of each feature. This baseline input helps to understand how the output changes relative to a standard or neutral state, allowing for a clearer interpretation of feature importance [24].

  The calculation of SHAP values in Deep SHAP involves several key steps:

  1. **Generating a Background Dataset**: Create a background data set that captures the distribution of the input features in the training data.
  2. **Sampling Reference Points**: Sample a subset of this background dataset to establish reference points against which the contributions of features will be evaluated.
  3. **Model Output Generation**: For each reference point, generate the model output to assess how changes in features influence predictions, thus allowing the calculation of SHAP values based on the differences between the model output for reference points and the actual input.

  This process helps to accurately attribute the influence of each feature in the context of complex interactions typical in deep learning architectures.

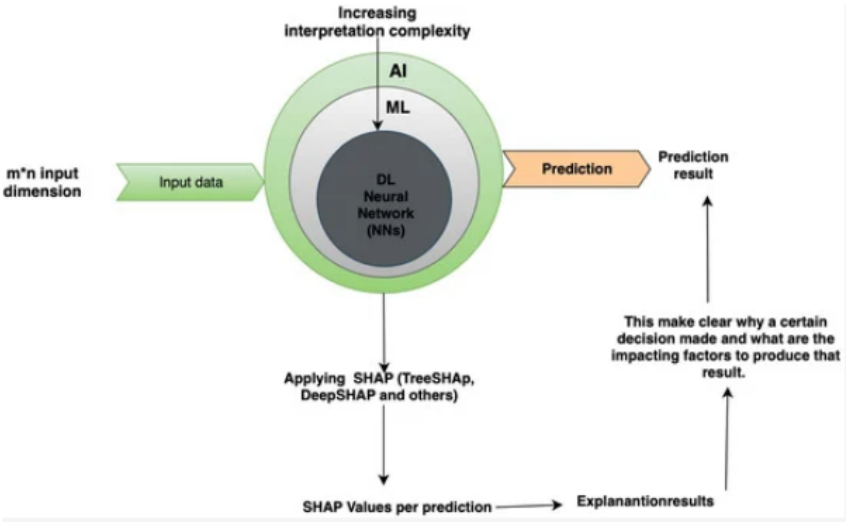

Figure 8: The SHAP framework applied to any DL model [8]

4 Proposed Root Cause Analysis Framework

This chapter presents a RCA framework designed to identify the underlying causes of high PUE rates in data centers. The framework integrates DL and SHAP to leverage advanced machine learning techniques for a deeper understanding of energy inefficiencies. The proposed approach aims to lower operating expenses, increase energy efficiency, and encourage sustainability in data center operations.

4.1 RCA Framework Overview

Data centers are critical infrastructures that consume substantial amounts of energy, making energy efficiency a priority for operational sustainability and cost management. PUE is a key performance indicator that is used to measure energy efficiency in data centers. It is defined as the ratio of total building energy usage to energy consumed by IT equipment alone. A lower PUE indicates better energy efficiency, while higher values suggest significant energy wastage, often due to inefficient cooling and other non-IT functions. High PUE rates mean higher operational costs and more energy consumption that has a negative impact on the environment.

In order to address these inefficiencies, this framework proposes a comprehensively enhanced plan that integrates DL and SHAP for the application of energy efficiency in energy-use strategy. Using DL to analyze the time series of data on energy consumption, the framework aims to identify significant patterns and anomalies in energy use. This analysis helps to understand how energy is consumed over time and pinpoints any irregularities that may indicate inefficiencies. Meanwhile, SHAP provides interpretive insights into the factors driving high PUE rates, allowing data center operators to understand and effectively address the root causes.

The integration of DL and SHAP offer multiple benefits. First, it helps reduce operational cost by identifying and addressing inefficiencies, leading to significant savings. Second, reducing PUE contributes to sustainability efforts by reducing energy waste, thus minimizing the environmental footprint of data centers. Lastly, the framework improves decision making by providing insight into recommendations that empower operators to implement effective energy management strategies.

In summary, the proposed RCA framework combines the strengths of DL and

SHAP to provide a robust solution to improve energy efficiency in data centers. Using a focus on identifying and addressing the root causes of high PUE rates, this innovative approach promotes sustainability and operational excellence.

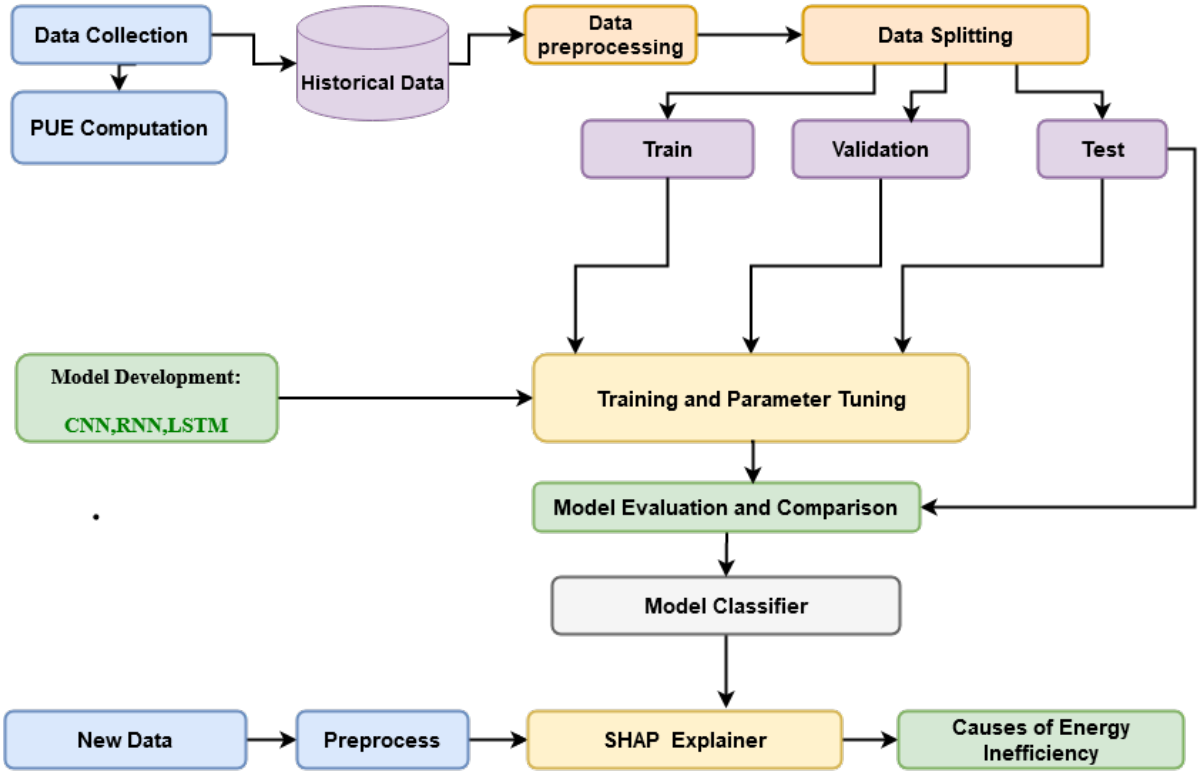

Figure 9: Present a proposed approach for high PUE rate RCA

4.2 Experimental Design

1.Data Collection

When performing an RCA of inefficient PUE, it is crucial to systematically identify and address the underlying problems that affect power consumption. This study focuses on analyzing power consumption equipment to determine the root causes of power consumption inefficiencies. To achieve this, data were collected from the EP&EMS of ethio telecom, which provides a comprehensive view of the patterns of energy use. The data collection process entails multiple important stages. To begin with, finding main PUE factors is very important; this also means studying and making a list of possible causes of high or inefficient PUE rates. These

reasons must be clearly defined and understandable to facilitate effective communication and action. The relevant parameters must then be pinpointed, focusing on critical metrics such as energy-related conditions and environmental conditions such as temperature and humidity. This targeted approach ensures that the most impactful factors that influence PUE are monitored. Subsequently, a series of data gathering tasks are performed. This includes real-time monitoring of power consumption in all relevant equipment and collecting environmental data to capture conditions within the data center.

Historical data is also analyzed to identify trends, while qualitative insights from surveys and interviews with data center staff help uncover operational practices that may contribute to inefficiencies and organizing the data effectively prepares it for analysis. To determine specific parameters that can be measured within the data center, directly linking them to identified causes of inefficiency. Key measurement parameters include grid total power, total input UPS power, and total output UPS, among others. Each of these parameters provides valuable information on the performance of PUE in data center. The data set comprises 24 features that serve as causal factors for PUE inefficiency, offering a comprehensive understanding of the operational environment and equipment performance within the data center. The following measurements serve as input parameters for the DL model, each contributing valuable insights into the energy efficiency dynamics of the data center.

Grid Power Parameters

- Grid Power Phases A, B, and C (unit:kw):These represent the power consumption of each of the three phases of the electrical grid. This is essential for understanding the distribution of power between phases and identifying any imbalances.
- Total power of the grid (unit:kw): This is the sum of the power consumption of the three phases, providing an overall picture of the power drawn from the grid.

UPS Power Parameters

- Input UPS power phases A, B, and C (unit:kw): These measure the power

input to the UPS from each phase of the grid. This helps assess the load on the UPS and identify any phase-specific issues.

- Total input UPS power (unit:kw): This is the total power input to the UPS from all phases.
- Output UPS power phases A, B, and C (unit:kw): These measure the power output from the UPS to each phase of the load. This indicates the power that is being delivered to the data center.
- Total output UPS power (unit:kw): This is the total power output of the UPS in all phases.
- UPS efficiency: This is a measure of how efficiently the UPS converts input power to output power. A higher efficiency rating indicates less power loss.
- UPS power loss: This is the power lost within the UPS system due to inefficiencies.

Rectifier Power Parameters

- Rectifier input power (unit:kw): This is the power input to the rectifier, which converts AC power to DC power.
- Rectifier output power: This is the power output from the rectifier that is delivered to the DC loads.
- Rectifier loss power: This is the power lost during the rectification process.

Load Power Parameters

- Air conditioner power (unit:kw): This is the power consumed by the air conditioning units to maintain the desired temperature and humidity in the data center.
- Access load power:(Unit:kw) This is the power consumed by non-critical loads, such as lighting, security systems, and other miscellaneous equipment.
- IT load power(unit:kw): This is the total power consumed by the IT equipment, calculated as the sum of the UPS output and the rectifier output power.

Environmental Parameters

- Outdoor temperature (unit: ° C): This is the ambient temperature outside the data center.
- Outdoor RH (unit:%): This is the relative humidity outside the data center.
- Temperature setting point (unit: ° C): This is the desired temperature setting for the data center environment.
- RH setting point (unit: ° C): This is the desired relative humidity setting for the data center.
- IT room temperature (unit: ° C):This is the actual temperature within the IT room.
- RH in the IT room: This is the actual relative humidity within the IT room.

Efficiency Metric

- PUE : This is a key metric that indicates the overall energy efficiency of the data center. It is calculated as the total power consumption of the facility divided by the IT load power. A lower value of PUE indicates a higher efficiency.

This data set not only provides a comprehensive overview of operational conditions,but also facilitates mapping of causes to symptoms, as illustrated in the accompanying table [2]. For example, high Air Conditioner Power may indicate inefficient cooling systems, leading to excessive energy consumption. Conversely, a high Outdoor Temperature can result in increased cooling demands, further straining the energy resources of the data center. Therefore, data center operators must analyze the collected data and implement appropriate measures to improve overall performance.

Table 2: Cause symptom mapping for PUE inefficiency

Input Parameter	**Potential Cause of High PUE**	**Symptom**
UPS Input Power (High)	High load on the data center	Increased likelihood of UPS reaching capacity
UPS Output Power (Low)	Inefficient UPS operation	Lower power availability for IT equipment
UPS Loss Power (High)	Inefficient UPS conversion	High energy waste within the UPS system
Rectifier Input Power (High)	High DC load or inefficient conversion	Increased energy consumption for DC loads
Rectifier Output Power (Low)	Inefficient rectifier conversion or low DC load	Lower power availability for DC equipment or potential rectifier issue
Rectifier Loss Power (High)	Inefficient AC to DC conversion	High energy waste during rectification
Air Conditioner Power (High)	Inefficient cooling system or improper temperature settings	Increased energy consumption for cooling
IT Load Power (Unexpectedly High)	Inefficient IT equipment, high server utilization, or hardware issues	Increased energy consumption for IT operations
Outdoor Temperature (High)	Increased cooling demand	Higher air conditioner power consumption
Outdoor Relative Humidity (High)	Reduced cooling effectiveness	Air conditioning system working harder to maintain humidity levels
Temperature Setpoint (Low)	Overly cold data center environment	Unnecessary cooling and higher energy consumption
IT Room Temperature (Higher than Setpoint)	Inefficient cooling or insufficient airflow	Potential equipment overheating and risk of failures
IT Room Relative Humidity (Higher than Setpoint)	Inefficient humidity control	Potential equipment damage due to excessive humidity

After data collection, the next critical step is the calculation of PUE. This process begins with understanding the power flow within the data center, which involves measuring the active power drawn from all electrical phases,phase A, phase B, and phase C. Accurate measurements taken from the main distribution board allow operators to capture the total energy consumed by both IT and non-IT components. Following this, Total Facility Power is calculated by summing the active power from each phase, encompassing all energy used by non-IT infrastructure, such as cooling systems and lighting.

The power of IT equipment is evaluated, focusing specifically on the energy consumed by the IT devices. This is determined by adding the rectifier output power and the UPS output power, which reflect the power used by the servers and storage devices. With both the total power facility and the power IT equipment calculated, the PUE can be calculated using the established formula. This calculation provides valuable information on the energy efficiency of the data center. In particular, the real measurement and quantitative findings of PUE did not appear in the tool until this section, emphasizing the importance of accurate computation for effective monitoring and management. By analyzing the resulting PUE value, operators can identify inefficiencies, optimize resource allocation, and implement strategies to reduce operational costs, ultimately supporting sustainability goals. In addition, historical data can be segmented by varying parameters, such as time of day,week, or seasonal factors. This granular analysis allows operators to understand how different conditions affect energy consumption and PUE.

2.Data Preparation

This is a crucial phase in the preparation of the collected data for analysis and modeling, ensuring that the data set is clean, consistent and suitable for machine learning algorithms. This step improves the accuracy and effectiveness of subsequent analyzes, allowing for more reliable predictions regarding PUE inefficiencies in the data center.

- **Data cleaning**: In data preprocessing, the initial step is data cleaning that requires to find and fix errors or inconsistencies in the dataset. This process includes removing irrelevant data, fixing formatting issues, and addressing missing values. Missing values can significantly skew results, so they are han-

dled through imputation or deletion, depending on their extent and impact on the data set. Additionally,any duplicate entries are removed to maintain data integrity. By ensuring that the data set is free from errors, operators can rely on the accuracy of their analyses.

- **Normalization and Scaling**: Once the data are cleaned, normalization is performed to bring all the features to a common scale, especially since different parameters may have varying units and ranges. This step is particularly important for machine learning algorithms, which can be sensitive to the scale of input features. Robust scaling techniques, such as using the interquartile range (IQR), can be used to mitigate the influence of outliers while ensuring that all features contribute equally to the performance of the model.
- **Feature engineering**: This is another vital aspect of data pre-processing. This involves creating new features that can enhance the predictive power of the model. For example, combining multiple related parameters or converting continuous data into categorical labels can provide additional information. By enriching the data set with relevant features, operators can improve the ability of the model to identify patterns associated with high PUE values.
- **Data discretization**: Is the process of converting continuous data into categorical data, which is particularly useful for classification tasks. In the context of PUE analysis, we establish a binary classification system for the PUE values:
  - **Class 0**: High PUE (values equal or above 2.5)
  - **Class 1**: Not high PUE (values at or below 2.5)

  This classification allows for a clearer understanding of energy efficiency within the data center. By categorizing PUE values in this way, we can facilitate the identification of data center that require attention and optimization. This binary approach simplifies analysis and helps to develop targeted strategies to improve energy performance.
- **Data exploration**:Is a crucial phase in understanding the structure and characteristics of a dataset before analysis and modeling. It involves generating summary statistics to assess central tendencies and dispersion, checking for null values to identify areas that need attention, and visualizing data distri-

butions through histograms and box plots to reveal skewness and outliers. Additionally, correlation analysis using heatmaps helps uncover relationships between features, guiding feature selection and engineering. In general, these techniques provide valuable insights that inform preprocessing and model development, ensuring a solid foundation for effective data analysis.

Finally, the pre-processed data are organized and divided into subsets for training, validation, and testing. This systematic split ensures that the machine learning models can be trained effectively while being evaluated against unseen data. By preparing the data set in this manner, operators set the stage for successful analysis and modeling, which ultimately leads to improved energy efficiency in the data center. Through these preprocessing steps, the data set is effectively prepared for analysis, maximizing the potential for accurate predictions regarding PUE inefficiencies and enabling data center managers to implement targeted strategies for improvement.

3.Tools and Software used

To implement RCA, several open source software tools were used, providing a robust environment for data analysis and machine learning.

- Python: Anaconda distribution: A custom Python distribution for data science, including essential packages:
  - NumPy: For scientific computing and handling multidimensional data.
  - Scikit-learn: For data analysis and machine learning tools.
  - Pandas: For efficient manipulation and processing of tabular data.
  - Keras and TensorFlow: For building and training neural networks.
  - SHAP (SHapley Additive exPlanations) Python library; is used for interpreting the output of machine learning models.

4.3 Development of the 1D CNN Model

To develop a 1D CNN for binary classification, our objective is to determine whether the performance of a data center is efficient (class 1) or inefficient (class 0). The model is implemented using Keras and follows a sequential architecture, which

allows us to stack layers in a linear fashion. This approach simplifies the model construction, making it easy to understand and modify as needed. The sequential model is particularly well-suited for this task, as it can efficiently process time series data or sequentially structured data common in performance analysis.

The architecture of the model consists of two convolutional layers, each containing 128 filters with a kernel size of 3, coupled with ReLU activation functions. In order for the model to learn spatial hierarchies, these convolutional layers are necessary for removing significant features from the input data. The choice of 128 filters enables the model to capture a rich set of features, enhancing its ability to distinguish between the two classes. Following the convolutional layers, a dropout layer is included with a rate of 0.5 to mitigate the risk of overfitting. This layer randomly sets a portion of the input units to zero during training, promoting generalization and improving the model's performance on unseen data. To further refine the model, a max-pooling layer is employed after the dropout layer. This layer reduces the dimensionality of the feature maps, summarizing the presence of features while retaining the most critical information. The pooling operation helps reduce the computational load and the number of parameters in the model, further reducing the risk of overfitting. After the pooling layer, a flatten layer converts the multi-dimensional output into a one-dimensional array, making it compatible with the subsequent dense layers. The model concludes with two dense hidden layers, each containing 64 units and utilizing ReLU activation functions. These layers enable the model to learn complex patterns and relationships within the data.

Finally, the output layer consists of a single unit with a sigmoid activation function, which is ideal for binary classification tasks. The model is compiled using the binary cross-entropy loss function, which assesses the model's performance by comparing the predicted probabilities to the true class labels. This structured approach ensures that the 1D CNN effectively learns to classify the performance of data center with accuracy.

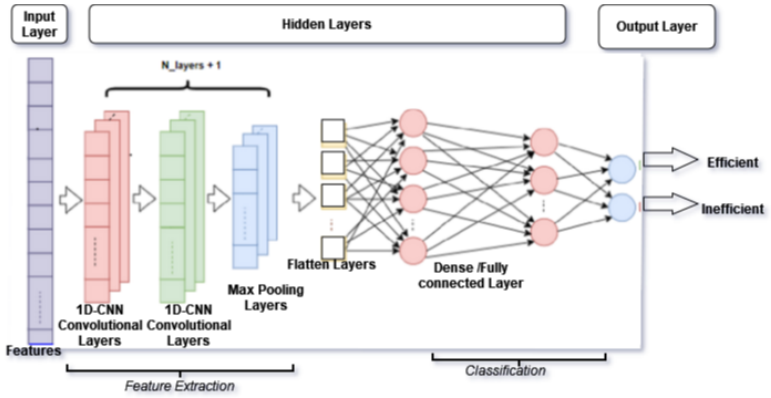

Figure 10: 1D CNN model architecture for data denter performance classification [9]

4.4 Evaluation of the Model

Evaluating the performance of a DL models in classifying PUE is crucial to understanding its effectiveness in real-world applications. A comprehensive assessment involves several key metrics that provide insight into different aspects of model performance.

1. Confusion matrix

The confusion matrix is a powerful tool to visualize the performance of the classification model. Displays the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in matrix format. This representation allows us to quickly identify how many instances were correctly or incorrectly classified. For example, a well-performing model will have high TP and TN values while minimizing FP and FN. The confusion matrix helps in deriving other performance metrics and provides a clear understanding of where the model may be making errors.

2. Sensitivity (recall)

Sensitivity, also known as recall, measures the model's ability to identify inefficient instances (class 0). It is calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.1}$$

A high sensitivity value indicates that the model is effective in capturing most of the inefficient , which is vital for operational efficiency and intervention strategies. This

metric is particularly important in scenarios where the lack of an efficient instance could lead to significant energy waste.

3. Precision

Precision assesses the accuracy of the model in classifying positive instances (class 0). It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

A high precision value means that when the model predicts an instance as inefficient, it is likely to be correct. This metric is crucial to minimize false alarms and ensure that resources are effectively allocated to only this data center that genuinely needs attention.

4. Specificity

Specificity evaluates the model's ability to correctly identify efficient instances (class 0). It is calculated as follows:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.3}$$

High specificity indicates that the model effectively recognizes data center that are performing efficiently,thus reducing the likelihood of unnecessary interventions. This balance between sensitivity and specificity is essential for effective decision-making.

5. F1 score

The F1 score provides a fair assessment of the model's performance by combining accuracy and recall into a single statistic. It is particularly useful when dealing with imbalanced datasets. The F1 score is calculated as follows:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

A high F1 score indicates that the model maintains a good balance between precision and recall, signifying its overall effectiveness in correctly classifying instances.

6. Accuracy

Accuracy measures the overall correctness of the model's predictions and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

Although accuracy is a straightforward metric, it can be misleading in unbalanced data sets where one class significantly outnumbers the other. Thus, it should be considered alongside the other metrics to obtain a comprehensive view of the model's performance. A comprehensive evaluation of the performance of the 1D CNN model in classifying PUE efficiency involves analyzing the confusion matrix, sensitivity, precision, specificity, F1 score, and accuracy. Each of these metrics provides valuable information on different aspects of model performance, ensuring that the model can effectively support decision making regarding data center efficiency and enhance strategies.

5 Results and Discussions

The performance of the machine learning model is essential for evaluating its accuracy and reliability in real-world applications. This evaluation will use a nine-month data set of 6,586 hourly power consumption measurements from a data center, focusing on classifying PUE efficiency as a key metric in data center management. By accurately classifying the efficiency of the PUE , the model will provide information on operational performance and energy use. The EP&EMS will facilitate a detailed analysis of the power consumption patterns.

Collaboration with the data center operator team will improve the understanding of operational dynamics, allowing the identification of areas of high consumption and potential problems. This approach will improve the credibility of the evaluation and support actionable energy management strategies.The findings will demonstrate the predictive capabilities of the model and its potential impact on energy efficiency and sustainability in the data center.

5.1 Dataset Visualization

The dataset used in this study is comprehensive and well structured, with no missing values, improving the reliability of the analysis and ensuring that all observations contribute to the training and evaluation of the model.

A. Visualization of Target Variables

In this data set, the PUE metric is analyzed with the following statistics:

- Mean PUE: The average PUE value across the data set.
- Standard Deviation: Indicates the variability of the PUE values.
- Range: The lowest and highest recorded PUE values.
- Quartiles: Breakdown of PUE values into four segments, including the median and interquartile range.

The analysis of these statistics provides information on the energy efficiency of the data center and helps identify patterns and trends that are critical for further investigation. The data set offers a comprehensive overview of the PUE metric, critical

evaluating the energy efficiency of the data center.

The visualization of this statistics (Figure 11) indicates a mean PUE of 2.504119, suggesting that on average data center consume 2.5 times more energy than the minimum required for the computing equipment. A relatively low standard deviation of 0.174057 implies that the PUE values show minimal variability, reflecting a consistent energy efficiency profile in the data centers analyzed.

Key Insights

- **PUE Range:** The lowest observed PUE is 2.20000, while the highest is 3.00000. This range indicates significant potential for improving energy efficiency, particularly as PUE values above 2.5 are generally considered inefficient.
- **Quartile Breakdown:**
  - 25th Percentile: 2.33000 (25% of data center at or below this value)
  - Median (50th Percentile): 2.53000 (half of the data center at or below this level)
  - 75th Percentile: 2.65000 (75% of data center at or below this threshold)

	PUE
count	6586.000000
mean	2.504119
std	0.174057
min	2.200000
25%	2.330000
50%	2.530000
75%	2.650000
max	3.000000

Figure 11: Statistical summary of PUE

These findings emphasize the ongoing need to monitor and improve energy efficiency in the data center. A significant portion of the facility's PUE values exceed

the optimal range, indicating clear opportunities for improvement. By adopting a data-driven approach, operators can identify specific areas that require targeted interventions. Implementing strategies based on these insights will help improve overall energy efficiency and reduce operational costs.

Figure 12 presents a histogram of the PUE distribution for the data center, with values ranging from 2.2 to 3.0. Most values are concentrated between 2.4 and 2.7, peaking at 2.5, indicating that the facility generally operates at a high efficiency level. However, the slight positive skew in the distribution suggests that there are instances of inefficient energy usage, particularly for PUE values exceeding 2.7.This variation in PUE highlights the need to investigate the root causes of energy inefficiency within the data center.

By analyzing the histogram, operators can gain insight into the specific areas of concern and develop targeted strategies to address the factors that contribute to inefficiency, ultimately enhancing energy management practices within the facility.

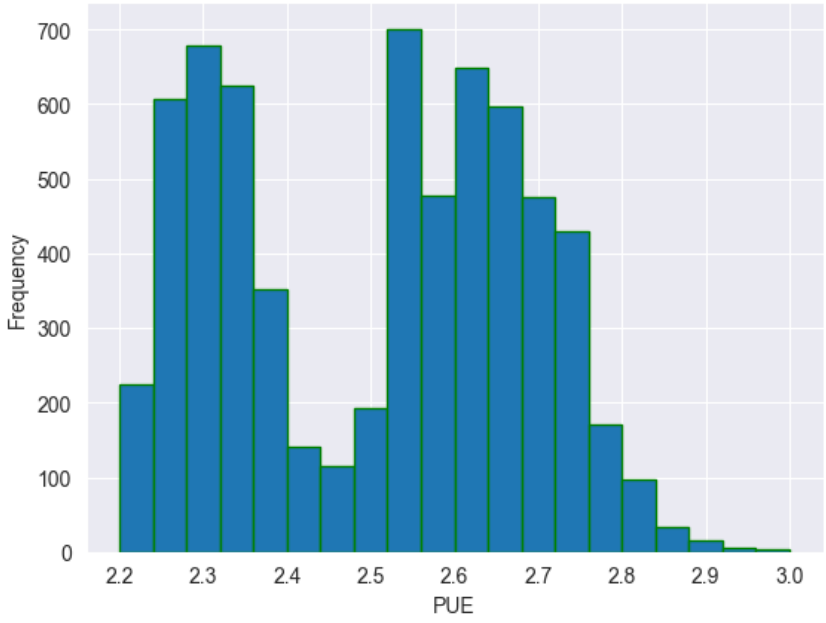

Figure 12: PUE distribution histogram

B. Feature correlation

Figure 13 illustrates the correlation heatmap, providing essential information on the relationships between various features and PUE.

- **Positive Correlation:** Features highlighted in green, indicating a strong positive correlation with PUE. Key variables, such as Air Conditioning Power (unit: kW), UPS loss power and IT load power (unit: kW), fall into this category. This positive correlation suggests that as the values of these in the characteristics , there is a corresponding increase in the PUE value. Essentially, a higher energy consumption from cooling systems and IT loads is associated with a less efficient overall energy use in the data center.
- **Negative Correlation:** In contrast, features displayed in brown show a strong negative correlation with PUE. For instance, Rectifier Efficiency (unit: %) and UPS Efficiency (unit: %) demonstrate this inverse relationship. As the efficiencies of these systems increase, the value of the PUE tends to decrease. This finding emphasizes the importance of improving the efficiency of power systems to improve overall energy performance.
- **Feature Selection:** To refine the analysis, several features with very low correlation to PUE were excluded from the analysis. These include variables such as Temp_set_point (unit: C), RH_set_point (unit: %), Out_Door_Temp (unit: C), Out_Door_RH (unit: %), UPS_Total_Input/Output_P (unit: kW), and Rectifier_Input/Output_P (unit: kW). This targeted approach allows for a more focused analysis, focusing on the most relevant features that significantly affect PUE.
- **Interpretation:** The correlation heatmap serves as a valuable tool for understanding the dynamics between input variables and PUE. The identified positive (green) and negative (brown) correlations can guide the feature selection process and inform the development of predictive models.
- **Further Analysis:** Looking ahead, the analysis will continue by focusing on highly correlated features to uncover the root causes of elevated PUE rates.

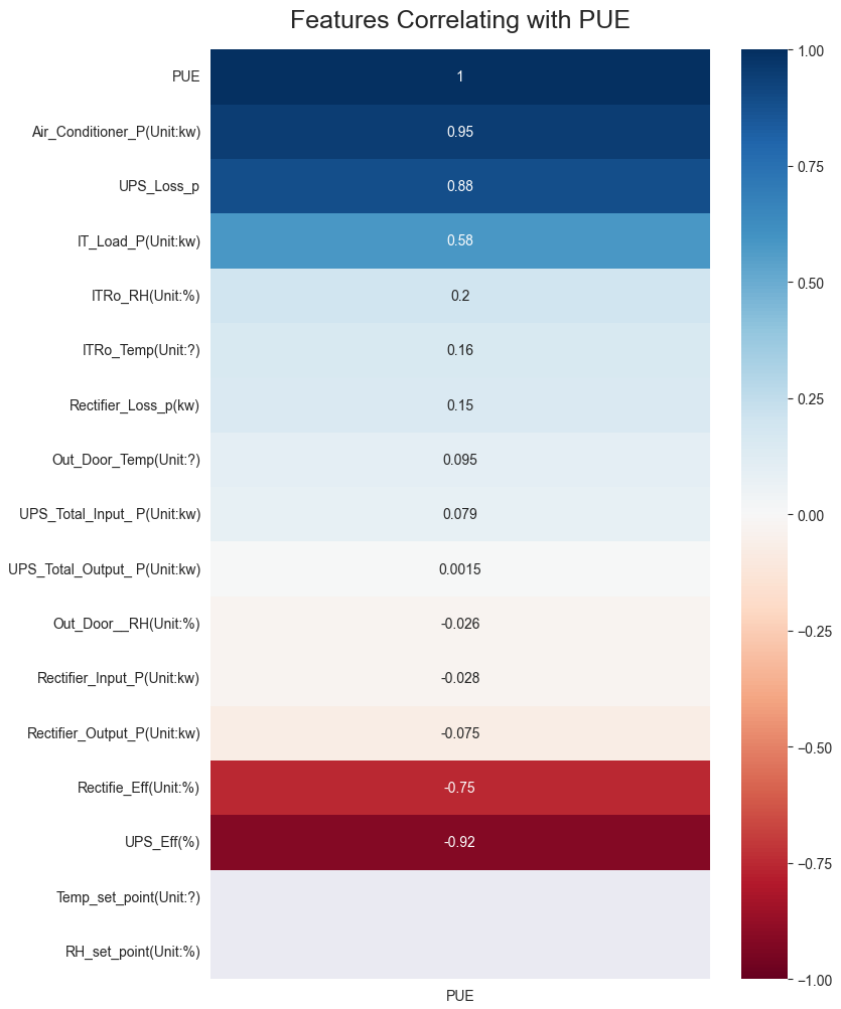

Figure 13: Correlation between features and the target

5.1.1 Comparative Analysis of Models

The model evaluation process compared three deep neural network architectures: 1D CNN, RNN and LSTM. Models are evaluated side by side to determine which architecture best suits the specific problem at hand. This analysis typically involves training different models with varying architectures, hyperparameters, or algorithms on the same data set. Using consistent evaluation metrics,such as accuracy, precision, recall, and F1 score, can objectively assess each model's performance.

In performing a comparative analysis, it is also important to consider factors beyond just performance metrics. Computational efficiency, training time, and the model's ability to generalize to unseen data are critical aspects that influence the choice of the final model. As shown in Figure 14, the 1D CNN model achieved superior performance in all evaluation metrics, with particularly strong results in sensitivity

(0.999) and F1 score (0.999). The LSTM followed closely behind, while the RNN demonstrated more modest performance.

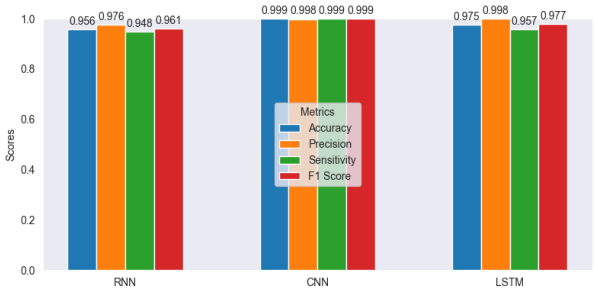

Figure 14: Model performance

The results show that for PUE classification, the 1D CNN pattern recognition capabilities outperformed the sequential modeling strengths of LSTM/RNN networks.

Table 3: Model performance comparison

Metric	**CNN**	**LSTM**	**RNN**
Accuracy	0.999	0.975	0.956
Precision	0.998	0.998	0.976
Sensitivity	0.999	0.957	0.948
F1 Score	0.999	0.977	0.961

5.2 Evaluation Outcomes

This section explores the performance of a DL model in a test dataset, highlighting the influence of hyperparameters during the training process. By analyzing the performance metrics of the model, we can gain insight into its effectiveness in classifying the data center based on their PUE. Understanding how hyperparameters impact the learning process is crucial to optimize model performance and improve prediction accuracy. Through rigorous evaluation, we can identify strengths and improve the model, ultimately informing strategies for energy efficiency in data centers.

5.2.1 Threshold Establishment for Classifying PUE

After analyzing the data, a threshold value was established for the binary classification of the PUE metric. Specifically, a PUE value less than 2.5 is classified as

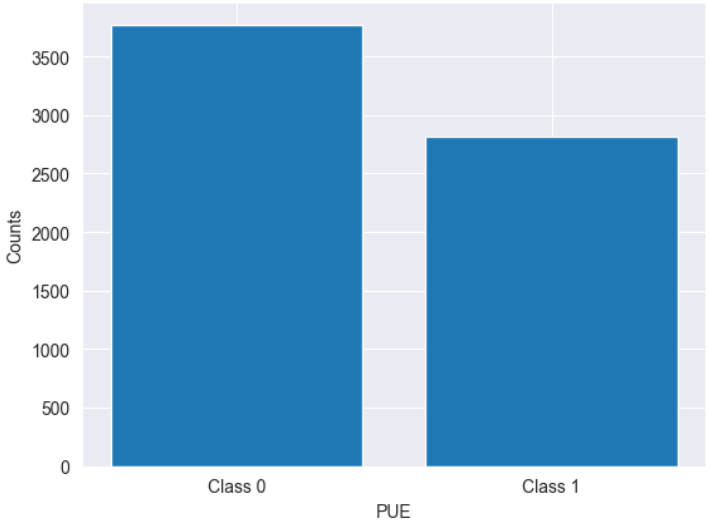

Figure 15: PUE classes

low or efficient (Class 1), while a value equal to 2.5 or greater is designated as a high or inefficient PUE rate (Class 0). This threshold was selected because it marks the point where PUE values tend to increase significantly.

By implementing this threshold, the model can effectively classify the data center as either having efficient or inefficient PUE. This binary classification approach is critical for identifying the key factors that contribute to high PUE rates, thereby forming strategies to improve energy efficiency in data centers.The choice of this threshold was guided by industry standards and expert recommendations, ensuring that the model's output aligns with recognized benchmarks for optimal PUE performance. This step is essential for defining the target variable and the corresponding classes, enabling the model to learn effectively and make accurate predictions.

5.2.2 Developed 1D CNN Model

The developed 1D CNN model is designed to classify the performance of the data center as efficient or inefficient based on PUE metrics. The following is a detailed overview of the model, including its architecture, training process, and evaluation results.

Model Architecture

The model architecture begins with an input layer that receives 18 input features

for analysis. The first Conv1D layer is designed to perform 1D convolution using 128 filters with a kernel size of 3 and a ReLU activation function. This configuration produces an output shape of (None, 18, 128) and contains 512 trainable parameters.

The second Conv1D layer mirrors the configuration of the first, producing an output shape of (None, 16, 128) and consisting of 49,280 trainable parameters. To mitigate overfitting, a dropout layer with a rate of 0.5 is incorporated.

Following this, a MaxPooling1D layer with a pool size of 2 is used to reduce the dimensionality of the feature maps. This layer has an output shape of (None, 8, 128) and contains 0 trainable parameters.

The flatten layer reshapes the tensor from (None, 8, 128) to (None, 1024), and the flattened output from these layers is then fed into two dense layers, each containing 64 neurons and utilizing ReLU activation, with a total of 65,600 trainable parameters.

Finally, the output layer consists of a single unit with a sigmoid activation function, appropriate for binary classification tasks, and has 65 trainable parameters.

Hyperparameter Tuning

The GridSearch method was utilized to fine-tune the hyperparameters of the model, allowing for systematic exploration of various configurations. This approach helps to identify optimal settings by evaluating multiple combinations of hyperparameters based on performance metrics. The hyperparameters that were tuned include several crucial aspects of the model's architecture and functionality:

- First, the number of convolutional layers was adjusted to determine the depth of the model, which can significantly impact its ability to learn complex patterns.
- In addition, the filter size was tuned, influencing how many features can be learned in each convolutional layer.
- The kernel size, which defines the size of the window that moves across the input data, was also considered, as it affects the granularity of feature extraction.
- Furthermore, the configuration of the dense layers and the number of neurons

within those layers were optimized, both of which play a critical role in the model's ability to capture intricate relationships within the data.

After hyperparameter tuning using GridSearch, the summary of the developed model for the dataset is presented in Figure 16.

```
model1=create_cnn_model()
model1.summary()
```

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_10 (Conv1D)          (None, 18, 128)           512

 conv1d_11 (Conv1D)          (None, 16, 128)           49280

 dropout_5 (Dropout)         (None, 16, 128)           0

 max_pooling1d_5 (MaxPooling  (None, 8, 128)           0
 1D)

 flatten_5 (Flatten)         (None, 1024)              0

 dense_11 (Dense)            (None, 64)                65600

 dense_12 (Dense)            (None, 1)                 65

=================================================================
Total params: 115,457
Trainable params: 115,457
Non-trainable params: 0
```

Figure 16: Summary of the proposed 1D CNN model

The trainable parameters,which include the weights and biases within the model,are updated throughout the training process. These parameters are vital because they are learned from the input data, enabling the model to make predictions and optimize its performance based on feedback from the loss function. Adjusting these parameters during training is essential for enhancing the model's accuracy and generalization capabilities, allowing it to better fit the encountered data.

This comprehensive overview of model development and architecture elucidates the key design choices and hyperparameters explored to optimize the performance of the binary classification model. The model is compiled with the binary cross-entropy loss function, the Adam optimizer, and the accuracy as the evaluation metric. The proposed DL classification hyperparameters are summarized in Table 4.These hyperparameter settings are crucial for the model's performance and reliability, reflecting careful selection and tuning to optimize the DL for binary classification tasks. The choices regarding the optimizer, learning rate, and loss function directly influence the model's learning process and its convergence towards an

optimal solution. Moreover, the cross-validation strategy bolsters the model's reliability, while the number of epochs and batch size regulate the training dynamics. The hyperparameter values provided indicate a thoughtful approach to achieving the best possible performance for the binary classification model.

Table 4: DL model overall hyperparameter settings

Parameter	**Value**
Optimizer	Adam
Learning Rate	0.001
Loss	Binary Cross-entropy
Stratified Cross Validation	Fold Size: 10
Epochs	50
Batch Size	64

5.2.3 Performance of the Developed 1D CNN Model

The performance of the developed model can be effectively summarized through the analysis of its confusion matrix. The confusion matrix is a key tool for evaluating the accuracy of a classification model by providing a detailed breakdown of its predictions. In this case, the confusion matrix presents the counts of true positives, true negatives, false positives, and false negatives, facilitating a comprehensive understanding of the model performance.

From the confusion matrix we observe that the model accurately predicted a significant number of instances for both classes. Specifically, it correctly classified 754.4 instances of the negative class (Actual 0) and 560.2 instances of the positive class (Actual 1). These values indicate a strong performance in identifying both classes, suggesting that the model effectively distinguishes between the two categories within the dataset.However, there are also instances of misclassifications that are critical to analyze. The matrix shows that there were 0.2 false positives, where the model incorrectly predicted a positive outcome for an actual negative instance, and only 2.4 false negatives, indicating a rare occurrence in which the model did not identify a positive instance. This low rate of false negatives is particularly encouraging, as it suggests that the model is adept at recognizing positive cases, which is

often crucial in binary classification tasks.

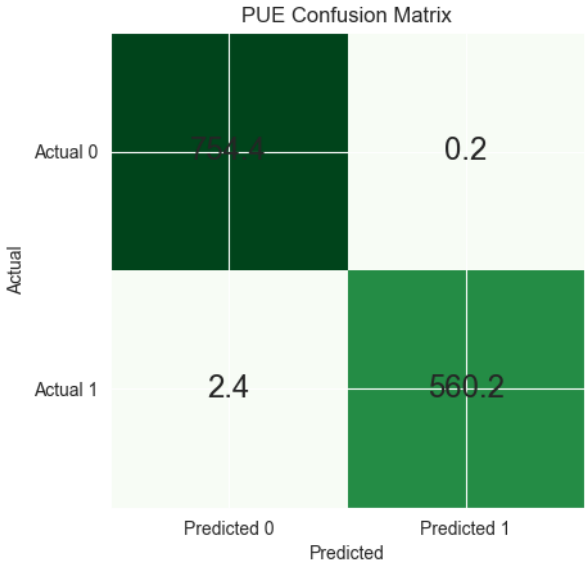

Figure 17: Model Performance Confusion Matrix Analysis

Furthermore, the following performance metrics further validate the effectiveness of the model.

Performance Metric	**Value**	**Description**
Accuracy	99.9%	Indicates the overall correctness of the model's predictions.
Precision	99.8%	Measures the ability to correctly identify positive instances.
Sensitivity (Recall)	99.9%	Shows the proportion of actual positive instances captured by the model.
F1 Score	99.9%	Balances precision and recall, reflecting overall reliability.

Table 5: Model Performance prediction

This comprehensive assessment highlights the robust performance of the model and shows its ability to minimize misclassifications while maintaining high accuracy across all performance metrics. This effectiveness is crucial for applications where distinguishing between classes is vital.

5.3 Results Overview

This section analyzes the input data for predicting high PUE rates (inefficiencies), employing both global and local interpretability methods. For global interpretability, SHAP values were visualized using the `shap.summary_plot()` function. This plot illustrates the impact of various features on model predictions, highlighting which inputs significantly influence the output, particularly within the identified inefficiency class. Visualization helps to understand the relationships between features and their overall contributions to model performance. Furthermore, a SHAP DeepExplainer object was created using a trained PUE model along with reshaped training data. This DeepExplainer computes SHAP values, indicating each feature's contribution to the model output. SHAP values were calculated for all samples in the reshaped testing dataset to provide a comprehensive view of the importance of the features.

For local interpretability, a KernelExplainer object was created to compute SHAP values for individual predictions. The the `shap.force_plot()` function was employed to generate a force plot, visually representing how each feature contributes to the predicted output for selected samples. This localized analysis provides information on specific instances, such as energy usage metrics of a particular data center on a specific day, allowing operators to identify critical areas for improvement.

By synthesizing insights from both global and local interpretability, this analysis offers a robust framework to understand the factors that drive high PUE rates, which allows targeted recommendations to improve operational efficiency in the data center.

5.3.1 Root Causes of High PUE Rates or Inefficiency Analysis

The analysis presents the most important features to predict whether the PUEis equal to or above a threshold in the test data, as illustrated in Figures 18 and 19. The model used for the prediction demonstrates a sensitivity of 99.9% and an F1 score of 99.9%.

Global Analysis of PUE Inefficiency Drivers

The results of the SHAP summary plot provides critical information on the factors that influence the inefficiency of PUE in the data center. This plot offers a comprehensive visualization of the impact different features have on model prediction. As shown in 18, the horizontal axis represents the mean absolute SHAP value, which measures the average magnitude of a feature's impact in all predictions, while the vertical axis lists the features in descending order of importance. Each dot represents an individual observation's feature contribution, with the color gradient (blue → red) encoding actual parameter values from low to high.

This visualization reveals that `Air_Conditioner_P(Unit.kw)` score emerges as the most influential predictor, with its mean SHAP value substantially exceeding all other features. The magnitude of this impact suggests that the cooling system is the dominant efficiency factor. Similarly, the `UPS_Loss_p` and `UPS_Eff(%)` features display notable variations, with some predictions indicating high losses, while others show more efficient performance. This variability highlights that optimizing uninterruptible power supply systems could mitigate inefficiencies, especially in instances with high UPS losses. Other features, such as `Rectifier_Loss_p(kw)` and `Rectifier_Eff(kw)`, exhibit a mix of low and high impact on the model's predictions, suggesting that rectifier systems are another crucial area for improvement.

Additionally, the `IT_Load(Unit.kw)` feature demonstrates a wide range of impacts on predictions, emphasizing the importance of effectively managing IT loads. Environmental factors, represented by `ITRo_Temp(Unit )` and `ITRo_RH(%)`, also show varying contributions, suggesting that maintaining optimal temperature and humidity levels is critical to reducing inefficiency.

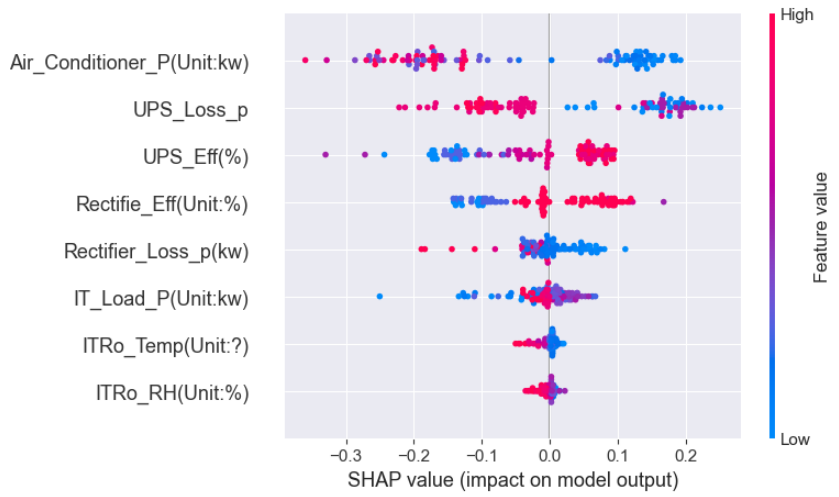

Figure 18: SHAP value distribution by feature: root causes of inefficiency

The accompanying bar graph, Figure 19 illustrates a bar graph generated to visualize the mean absolute SHAP values for each feature, showing their average impact on model predictions. SHAP values are calculated for the first 100 samples of the test dataset.

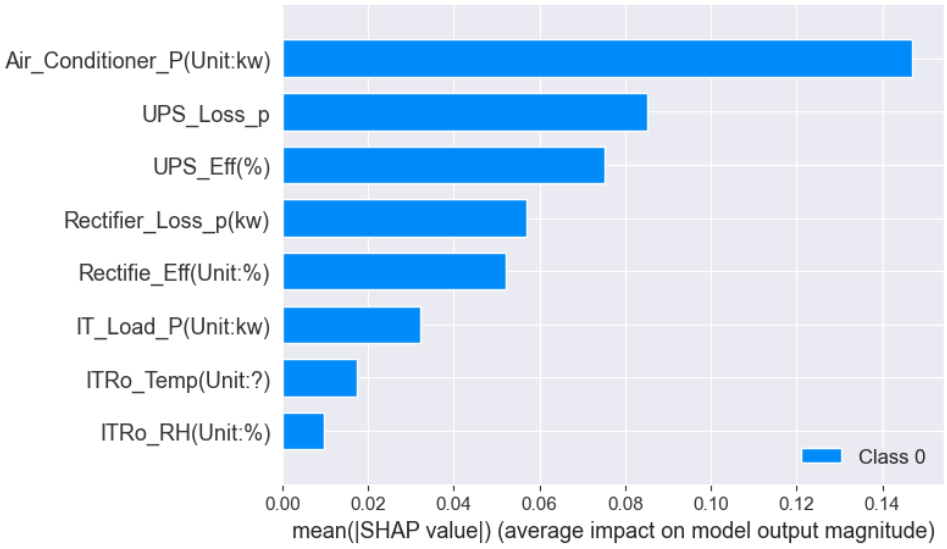

Figure 19: RCA of the average impact on PUE inefficiency for first 100 samples

Table 6: SHAP values and percentage contributions

Feature	SHAP Value	Percentage Contribution
Air_Conditioner_P(Unit: kW)	0.15	$\frac{0.15}{0.49} \times 100 \approx 30.61\%$
UPS_Loss_p	0.09	$\frac{0.09}{0.49} \times 100 \approx 18.37\%$
UPS_Eff	0.07	$\frac{0.07}{0.49} \times 100 \approx 14.29\%$
Rectifier_Loss_P(kw)	0.06	$\frac{0.06}{0.49} \times 100 \approx 12.24\%$
Rectifier_Eff	0.05	$\frac{0.05}{0.49} \times 100 \approx 10.20\%$
IT_Load_P(Unit: kW)	0.02	$\frac{0.02}{0.49} \times 100 \approx 4.08\%$
IT_Ro_Temp(Unit: °C)	0.01	$\frac{0.01}{0.49} \times 100 \approx 2.04\%$
Total	0.49	100.00%

The analysis identifies several root causes for the high PUE rates (inefficiency) in Class 0. The most significant contributor is `Air_Conditioner_P(Unit.kw)`, indicating excessive energy consumption by air conditioning systems. Additionally, `UPS_Loss_p` and `UPS_Eff(%)` highlight critical inefficiencies in uninterruptible power supplies, suggesting that improving their performance could reduce energy waste. `Rectifier_Eff(kw)` and `Rectifier_Loss_p(kw)` points to the need for optimization in rectifier systems, as their losses significantly impact overall energy consumption. Furthermore, `IT_Load(Unit.kw)` indicates that high-power loads from IT equipment are pertinent, emphasizing the importance of effective load management.

Lastly, environmental factors such as `ITRo_Temp(Unit )` and `ITRo_RH(%)` suggest that better control of temperature and humidity within facilities is essential for reducing inefficiency. Together, these factors highlight key areas to target to improve energy efficiency and reduce PUE rates.

Localized Analysis of Specific Inefficiency Cases

The force plot serves as an insightful visualization tool that allows one to examine individual instances and their contributions to the model output. This method enhances the interpretability of our analysis by clearly displaying the positive and negative impacts of each feature on the final prediction, relative to a baseline value. In the force plot, the length of each bar indicates the strength of a feature's influence on the model score: longer bars represent a more substantial impact, either positive or negative, while shorter bars suggest a minimal effect on the prediction.

This visualization provides operators with a clear view of how each feature in-

fluences the model output for specific instances, effectively identifying which factors elevate or lower the prediction. The base value acts as a reference point, typically reflecting the average prediction across the data set when no features are considered. The use of colors, red for positive contributions and blue for negative ones, further clarifies the contributions of each feature, illustrating how the model transitions from the base value to the actual prediction for the selected instance. This clarity enables data center operators to pinpoint the key drivers of PUE inefficiency, empowering them to make informed and targeted decisions to manage energy use.

In this analysis, we focus on calculating the SHAP values for a specific test sample, namely the 20th sample of the dataset. We start by reshaping this sample into a 1D array using the `flatten()` method, which is essential for effective visualization. Since we are dealing with binary classification, we extract the SHAP values relevant to class 0 (inefficiency) from the computed results. We then generate a force plot using the `shap.force_plot` function, which visually shows how each feature influences the model prediction for this sample.

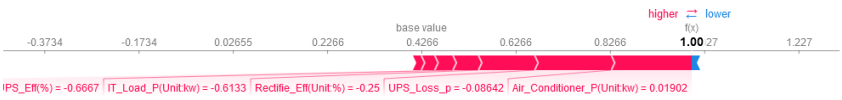

Figure 20: RCA of high PUE (inefficiency) in a single observation

For this instance, the base value of 1.00, indicating a high PUE rate, highlights significant contributors such as `Air_Conditioner_P(unit.kw)`, `UPS_Loss_p`, `Rectifier_Eff (unit:%)`,`IT_Load(unit.kw)`, and `UPS_Eff(unit:%)`, all of which positively affect the prediction. In contrast, `ITRo_Temp(unit?)` has a negative impact. This visualization is a powerful tool for understanding the model's decision-making process and the influential role of individual features in shaping the final prediction.

5.3.2 Discussion of the Result

The results of the SHAP analysis provide significant insight into the factors that contribute to the PUE inefficiencies in the data center. Using global interpretabil-

ity methods, such as `shap.summary_plot()`, provides a clear visualization of the contributions different features make to model predictions. The dominant role of air conditioning efficiency stands out, with a contribution of 30.61%, indicating substantial energy savings can be achieved by optimizing this critical component. Inefficiencies in air conditioning systems often stem from outdated equipment and insufficient environmental controls. Therefore, maintaining optimal temperature and humidity levels is essential not only for improving energy efficiency but also for ensuring the reliable operation of IT equipment. By adopting advanced cooling technologies and integrating real-time monitoring, data centers can significantly enhance cooling efficiency, thereby reducing PUE and improving overall operational effectiveness.

Additionally, the analysis underscores the importance of the UPS system, revealing that energy losses during power conversion and distribution account for 18.37% of inefficiencies. Furthermore, UPS efficiency contributes 14.29%, emphasizing that focusing on UPS performance can help mitigate these losses, contributing to a lower PUE. The findings also highlight inefficiencies in power conversion, as indicated by features such as `Rectifier_Loss_p(kw)` with a contribution of 12.24% and `Rectifier_Eff(Unit: %)` with 10.20%, which are critical to address. Upgrading to more efficient rectifiers and implementing advanced monitoring solutions can help identify and correct these inefficiencies, ensuring that the energy wasted in power conversion does not compromise the overall effectiveness of the data center.

Although the influence of IT equipment and environmental factors on PUE may be less pronounced, these elements remain significant, contributing 4.08% and 2.04% respectively. Conditions such as temperature, load, and humidity can greatly affect energy efficiency and operational reliability. Keeping optimal conditions for IT equipment is vital for achieving both high performance and energy efficiency. Continuous monitoring and adjustment of these environmental parameters can help ensure that data centers operate within their ideal ranges, further supporting efforts to reduce PUE.

The localized analysis provided by the force plot enhances understanding by illustrating how specific observations affect the model output. This detailed perspective allows operators to identify particular instances of inefficiency, facilitating

targeted interventions and helping to understand which features had the most significant impact on the predicted output. In the analysis for an individual instance, features such as Air Conditioner P(unit.kw) have a high impact on the prediction that are the root causes. In summary, the integration of global and local interpretability methods offers a robust framework for understanding the factors driving high PUE rates, empowering informed decision-making to improve energy efficiency in the data center.

6 Conclusion and Future Work

This chapter reviews the main results of the thesis and suggests ideas for future research. It also highlights the limitations of this study and discusses how these issues could be improved in future work.

6.1 Conclusion

The study successfully established a machine learning framework that combines a 1D Convolutional Neural Network (CNN) with SHapley Additive exPlanations (SHAP) to identify the root causes of energy inefficiency in the Nefas Silk Data Center. The 1D CNN model achieved outstanding performance, with accuracy, sensitivity, and F1 scores that all reached 99. 99%, surpassing other architectures like LSTM and RNN. By analyzing 6,586 hourly measurements, the framework pinpointed air conditioning systems as the main contributor to inefficiency, accounting for 30.61%, followed by power loss of the UPS at 18.37% and inefficiency of the UPS 14.29% and loss and inefficiencies of the rectifier performance at 12.24% and 10.20%. SHAP enhanced the model's interpretability, providing actionable insights into the factors affecting Power Usage Effectiveness (PUE).

Furthermore, the CNN-SHAP approach offers significant adaptability, allowing it to be tailored for various data Performance by simply adjusting the target variable for different data center configurations. Its extensive feature capacity enables the analysis of numerous features, facilitating detailed performance evaluations. Additionally, the framework provides temporal granularity, allowing for the identification of inefficiencies at an hourly level, which is more precise than traditional monthly or quarterly methods. These findings emphasize the transformative potential of machine learning in data center energy management , presenting scalable solutions that reduce energy waste, lower operational costs, and improve sustainability.

6.2 Future Work

To advance the proposed framework for enhancing energy efficiency in data centers, several key areas of focus have been identified:

- Improve RCA: The first priority is to enhance root cause analysis (RCA) by integrating historical data from multiple data centers. This integration will facilitate the identification of trends and patterns that can inform energy efficiency strategies in various facilities.
- Conduct Comparative Studies: Another crucial aspect involves conducting comparative studies in different geographical locations. This research will assess how regional factors, such as climate conditions, energy sources, and infrastructure, influence data center performance. Understanding these variations will enable the development of more tailored and effective energy management practices.

These future efforts aim to create a more robust and adaptable framework that not only addresses individual data center inefficiencies, but also promotes universal strategies for energy sustainability in diverse settings.

References

[1] C. Korra, "Sustainable design of data centers: A multidisciplinary approach," *International Journal of Open Publication and Exploration (IJOPE)*, 2024.

[2] K. Abiodun, "Digital infrastructure sustainable data centers investment in Africa: Role of Tier III Tier IV," 2025.

[3] N. L. Hanus, A. Newkirk, and H. Stratton, "Organizational and psychological measures for data center energy efficiency: barriers and mitigation strategies," *Energy Efficiency*, vol. 16, pp. 1–18, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255645194

[4] Y. Ran, X. Zhou, H. Hu, and Y. Wen, "Optimizing data center energy efficiency via event-driven deep reinforcement learning," *IEEE Transactions on Services Computing*, vol. 16, pp. 1296–1309, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:247305194

[5] Q. Fang, J. Zhou, S. Wang, and Y. Wang, "Control-oriented modeling and optimization for the temperature and airflow management in an air-cooled datacenter," *Neural Computing and Applications*, vol. 34, no. 7, pp. 5225–5240, 2022.

[6] M. N. Amin, M. Iqbal, M. Ashfaq, B. A. Salami, K. Khan, M. I. Faraz, and F. E. Jalal, "Prediction of strength and CBR characteristics of chemically stabilized coal gangue: ANN and random forest tree approach," *Materials*, vol. 15, no. 12, p. 4330, 2022.

[7] M. S. Ahmed and A. M. Fakhrudeen, "Deep learning-based COVID-19 detection: State-of-the-art in research," *International Journal of Nonlinear Analysis and Applications*, vol. 14, no. 1, pp. 1939–1962, 2023.

[8] Y. Gebreyesus, D. Dalton, D. De Chiara, M. Chinnici, and A. Chinnici, "AI for automating data center operations: Model explainability in the data centre context using Shapley additive explanations (SHAP)," *Electronics*, vol. 13, no. 9, p. 1628, 2024.

[9] M. Gheorghe, S. Mihalache, and D. Burileanu, "Using deep neural networks for detecting depression from speech," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, September 2023, pp. 411–415.

[10] P. K. Samuel, P. M. Moses, S. Davies, and C. Wekesa, "Analysis of energy utilization metrics as a measure of energy efficiency in data centres: Case study of Wananchi Group (Kenya) Limited data centre," in *2022 IEEE PES/IAS PowerAfrica*. IEEE, 2022, pp. 1–5.

[11] T. Zerihun, "Data center energy inefficiency root cause and sensitivity analysis: The case of Ethio-Telecom Legehar data center," Addis Ababa University, 2021, p. 70. [Online]. Available: http://etd.aau.edu.et/handle/123456789/29886

[12] L. Chen, X. Li, K. Wang, and X. Yu, "Optimization of energy efficiency of data center cooling systems," in *2024 36th Chinese Control and Decision Conference (CCDC)*, 2024, pp. 76–81. [Online]. Available: https://api.semanticscholar.org/CorpusID:271246490

[13] R. Kumar, S. K. Khatri, and M. J. Diván, "Power usage effectiveness (PUE) optimization with counterpointing machine learning techniques for data center temperatures," *International Journal of Mathematical, Engineering and Management Sciences*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244897857

[14] M. H. Jahangir, R. Mokhtari, and S. A. Mousavi, "Performance evaluation and financial analysis of applying hybrid renewable systems in cooling unit of data centers – a case study," *Sustainable Energy Technologies and Assessments*, vol. 46, p. 101220, 2021.

[15] S. Clement, K. Burdett, N. Rteil, A. Wynne, and R. Kenny, "Is hot IT a false economy? An analysis of server and data center energy efficiency as temperatures rise," *IEEE Transactions on Sustainable Computing*, vol. 9, pp. 482–493, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:265479675

[16] Y. Chen, K. Shi, M. Chen, and D. Xu, "Data center power supply systems: from grid edge to point-of-load," *IEEE Journal of Emerging and Selected Top-*

ics in Power Electronics, vol. 11, no. 3, pp. 2441–2456, 2022.

[17] M. Sharma, K. Arunachalam, and D. Sharma, "Analyzing the data center efficiency by using PUE to make data centers more energy efficient by reducing the electrical consumption and exploring new strategies," *Procedia Computer Science*, vol. 48, pp. 142–148, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:60215612

[18] M. Al-Amidie and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, pp. 1–74, 2021.

[19] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611–629, 2018.

[20] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[21] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[22] K. Papageorgiou, T. Theodosiou, A. Rapti, E. I. Papageorgiou, N. Dimitriou, D. Tzovaras, and G. Margetis, "A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing," *Frontiers in Manufacturing Technology*, vol. 2, p. 972712, 2022.

[23] T. Huang, T. Zhai, X. Zhang, and X. Di, "Electric power-grid friendly characteristic data center energy consumption optimization method," *Journal of Physics: Conference Series*, vol. 2095, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244486024

[24] G. Wassie, J. Ding, and Y. Wondie, "Traffic prediction in SDN for explainable QoS using deep learning approach," *Scientific Reports*, vol. 13, no. 1, p. 20607, 2023.

[25] M. Marcinkevics, A. Avotins, P. Apse-Apsitis, and A. Senfelds, "Identifying ineffective cooling implementation to increase energy efficiency in existing data

centres," in *2023 IEEE 10th Jubilee Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*. IEEE, 2023, pp. 1–6.

[26] J. C. Backhus and Y. Kono, "Incremental change detection method for data center power efficiency metrics (work in progress paper)," in *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257914191

[27] K. Sathupadi, "Deep learning for cloud cluster management: Classifying and optimizing cloud clusters to improve data center scalability and efficiency," *Journal of Big-Data Analytics and Cloud Computing*, vol. 6, no. 2, pp. 33–49, 2021.

[28] J. F. Baskoro, F. S. de Carvalho, and C. Apriono, "Root causes prediction in data center using convolutional dense neural network," in *2024 10th International Conference on Smart Computing and Communication (ICSCC)*. IEEE, 2024, pp. 7–10.

[29] H. L. Leka, Z. Fengli, A. T. Kenea, A. T. Tegene, P. Atandoh, and N. W. Hundera, "A hybrid CNN-LSTM model for virtual machine workload forecasting in cloud data center," in *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2021, pp. 474–478.

[30] W. P. Turner IV, J. PE, P. Seader, and K. Brill, "Tier classification define site infrastructure performance," Uptime Institute, vol. 17, 2006.

[31] Cisco, "Cisco global cloud index: Forecast and methodology, 2018-2023," 2020. [Online]. Available: [link] (Accessed: 2025-06-12).

[32] S. Cai and Z. Gou, "Towards energy-efficient data centers: A comprehensive review of passive and active cooling strategies," *Energy and Built Environment*, 2024.

[33] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, and M. Chinnici, "Machine learning for data center optimizations: feature selection using Shapley additive explanation (SHAP)," *Future Internet*, vol. 15, no. 3, p. 88, 2023.

[34] X. Shao, Z. Zhang, P. Song, Y. Feng, and X. Wang, "A review of energy efficiency evaluation metrics for data centers," *Energy and Buildings*, vol. 271, p. 112308, 2022.

[35] G. Ramezan, A. Abdelnasser, and Y. Ganjali, "KnowMe: A module to improve the efficiency of resource allocation in data center networks," in *2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2022, pp. 176–183. [Online]. Available: https://api.semanticscholar.org/CorpusID:247612434

[36] H. An and X. Ma, "Dynamic coupling real-time energy consumption modeling for data centers," *Energy Reports*, vol. 8, pp. 1184–1192, 2022.

[37] Q. Zhang, Z. Meng, X. Hong, Y. Zhan, J. Liu, J. Dong, T. Bai, J. Niu, and M. J. Deen, "A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization," *Journal of Systems Architecture*, vol. 119, p. 102253, 2021.

[38] M. Manganelli, A. Soldati, L. Martirano, and S. Ramakrishna, "Strategies for improving the sustainability of data centers via energy mix, energy conservation, and circular energy," *Sustainability*, vol. 13, no. 11, p. 6114, 2021.

[39] M. Wiboonrat, "Energy management in data centers from design to operations and maintenance," in *2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE)*, 2020, pp. 1–7. [Online]. Available: https://api.semanticscholar.org/CorpusID:231616349

[40] J. Cho, B. Park, and Y. Jeong, "Thermal performance evaluation of a data center cooling system under fault conditions," *Energies*, vol. 12, no. 15, p. 2996, 2019.

[41] M. Koot and F. Wijnhoven, "Usage impact on data center electricity needs: A system dynamic forecasting model," *Applied Energy*, vol. 291, p. 116798, 2021.

[42] C. Jin, X. Bai, C. Yang, W. Mao, and X. Xu, "A review of power consumption models of servers in data centers," *Applied Energy*, vol. 265, p. 114806, 2020.

[43] A. Bhattacharya, *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022.

[44] A. Katal, S. Dahiya, and T. Choudhury, "Energy efficiency in cloud computing data centers: a survey on software technologies," *Cluster Computing*, vol. 26, no. 3, pp. 1845–1875, 2023.

[45] J. Miguel-Alonso, "A research review of OpenFlow for datacenter networking," *IEEE Access*, vol. 11, pp. 770–786, 2022.

[46] P. Sun, Z. Guo, S. Liu, J. Lan, J. Wang, and Y. Hu, "SmartFCT: Improving power-efficiency for data center networks with deep reinforcement learning," *Computer Networks*, vol. 179, p. 107255, 2020.

[47] O. Van Geet and D. Sickinger, "Best practices guide for energy-efficient data center design," National Renewable Energy Laboratory (NREL), Golden, CO (United States), Tech. Rep., 2024.

[48] X. Liu, D. Teng, D. Wang, Q. Zhu, and Z. Liu, "Application of eco mode UPS in data center," in *2017 IEEE International Telecommunications Energy Conference (INTELEC)*. IEEE, 2017, pp. 30–34.

[49] Author Name, "The Revolution in the UPS Technology," ResearchGate, p. 374455127, 2023. [Online]. Available: https://www.researchgate.net/publication/374455127$_T he_R evolution_i n_t he_U PS_T echnology(A$ $2025 - 06 - 12$).

[50] Huawei, "UPS 2V battery rack installation guide," 2023. [Online]. Available: https://support.huawei.com/enterprise/ru/doc/EDOC1000034924?idPath=258788305

[51] ——, "Dual-live grid standalone co power system TP481200B-N20B2 datasheet," 2023. [Online]. Available: file:///C:/Users/Local

[52] E. E. Sadewa and S. Beta, "Temperature, humidity, and power outage monitoring system of Pamapersada Nusantara's server racks," *JAICT*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:256344410

[53] J. Liu, T. Wang, A. Skidmore, Y. Sun, P. Jia, and K. Zhang, "Integrated 1D, 2D, and 3D CNNs enable robust and efficient land cover classification from hyperspectral imagery," *Remote Sensing*, vol. 15, no. 19, p. 4797, 2023.

[54] Z. Li, H. Luo, Y. Jiang, H. Liu, L. Xu, K. Cao, H. Wu, P. Gao, and H. Liu, "Comprehensive review and future prospects on chip-scale thermal management: Core of data center's thermal management," *Applied Thermal Engineering*, p. 123612, 2024.

[55] J. Zou, Y. Han, and S.-S. So, "Overview of artificial neural networks," in *Artificial Neural Networks: Methods and Applications*, pp. 14–22. Springer, 2009.

[56] S. R. A. Parisineni and M. Pal, "Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations," *International Journal of Data Science and Analytics*, vol. 18, no. 4, pp. 457–466, 2024.

[57] S. Malik, M. Tahir, M. Sardaraz, and A. Alourani, "A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques," *Applied Sciences*, vol. 12, no. 4, p. 2160, 2022.

[58] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, pp. 1–74, 2021.

[59] M. Enkhbaatar and T. Yamazaki, "Automatic visual monitoring system for data center device management," in *2022 5th World Symposium on Communication Engineering (WSCE)*, 2022, pp. 22–25. [Online]. Available: https://api.semanticscholar.org/CorpusID:253122361

[60] X. Xue and N. Calabretta, "Nanosecond optical switching and control system for data center networks," *Nature Communications*, vol. 13, no. 1, p. 2257, 2022.

[61] Y. Chen, H. Xie, M. Ma, Y. Kang, X. Gao, L. Shi, Y. Cao, X. Gao, H. Fan, M. Wen et al., "Automatic root cause analysis via large language models for cloud incidents," in *Proceedings of the Nineteenth European Conference on Computer Systems*, 2024, pp. 674–688.

[62] T. Aditiyawarman, J. W. Soedarsono, A. P. S. Kaban, H. Rahmadani, R. Riastuti, and others, "Integrating the root cause analysis to machine learning interpretation for predicting future failure," *Heliyon*, vol. 9, no. 6, 2023.

[63] R. Kumar, S. K. Khatri, and M. J. Diván, "Effect of cooling systems on the energy efficiency of data centers: Machine learning optimization," in *2020 International Conference on Computational Performance Evaluation (ComPE)*. IEEE, 2020, pp. 596–600.

[64] J. N. Chukwunweike, C. C. Eze, I. Abubakar, L. O. Izekor, and A. A. Adeniran, "Integrating deep learning, MATLAB, and advanced CAD for predictive root cause analysis

in PLC systems: A multi-tool approach to enhancing industrial automation and fault diagnostics."

[65] S.-H. Han, K. W. Kim, S. Kim, and Y. C. Youn, "Artificial neural network: understanding the basic concepts without mathematics," *Dementia and Neurocognitive Disorders*, vol. 17, no. 3, pp. 83–89, 2018.