CNN-based Server State Monitoring and Fault Diagnosis using Infrared Thermal Images

Beltus Wiysobunri Nkwawir (nkwawir18@itu.edu.tr)
Istanbul Technical University https://orcid.org/0000-0002-9050-4055
Hamza Salih Erden
Istanbul Technical University - Ayazaga Campus: Istanbul Teknik Universitesi
Behcet Ugur Toreyin
Istanbul Technical University - Ayazaga Campus: Istanbul Teknik Universitesi

Research Article

Keywords: convolutional neural network, Infrared thermography, data center, servers, fault diagnosis

Posted Date: May 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1211668/v1

CNN-based Server State Monitoring and Fault Diagnosis using Infrared Thermal Images

Beltus Nkwawir Wiysobunri[1*], Hamza Salih Erden[1] and Behcet Ugur Toreyin[1]

[*]Informatics Institute, Istanbul Technical University, Maslak, Sariyer, 34467, Istanbul, Turkey.

*Corresponding author(s). E-mail(s): nkwawir18@itu.edu.tr;
Contributing authors: erdenh@itu.edu.tr; toreyin@itu.edu.tr;

Abstract

The recent spike in the demand for high-performance computing (HPC) server systems has created many challenges in data centers (DCs) including thermal management, system reliability sustenance and server failure minimalization. Lately, deep neural networks applied to infrared thermography (IRT) images have been successfully used for fault diagnosis in several fields. This paper evaluates seven state-of-the-art deep pretrained convolutional neural network (CNN)-based architectures and two shallow CNN-based architectures applied on server surface IRT images for the automatic diagnosis of five server operation conditions: partial CPU load; maximum CPU load; main fan failure; CPU fan failure; and server entrance blockage. Our approach is based on the concept of transfer learning which involves two main stages. First, a CNN model classifier pretrained on the large ImageNet dataset is used to extract lower level features. Second, the IRT images are used to fine-tune the higher levels of the CNN model classifier. A stratified five-fold cross-validation resampling method is used to evaluate the effectiveness and generalization of the nine architectures for five dataset split ratios. Results suggest that the CNN architectures achieve high prediction performance accuracies, with the majority having above 98% test accuracies across multiple split ratios. In addition, our diagnostic results are significantly higher than those obtained using a traditional support vector machine classifier trained on handcrafted features. The effectiveness and robustness of the CNN-based algorithms can provide

DC operators with an alternative intelligent approach to improve thermal management, energy efficiency, and system reliability of servers in DCs.

Keywords: convolutional neural network, Infrared thermography, data center, servers, fault diagnosis

1 Introduction

Over the past few years, data centers (DCs) have rapidly evolved to become the backbone of some of the world's most critical and prominent institutions such as banking, health, information, and communication technology (ICT) industries, etc. This fast-paced evolution and growth is fueled by the continuous rise in cloud computing and its related applications such as Big Data, Internet of Things, and Artificial Intelligence.

As a consequence, more high-density server systems have been developed that dissipate up to 30 kW per rack [1]. Today, some datacom facilities experience high overall power density loads in excess of $150W/ft^2$ [2]. This increase in server rack heat dissipation has given birth to new challenges in DC thermal management. For example, the uneven distribution of heat at the surface of servers and the lack of an efficient cooling system can result in the formation of localized hotspots around and inside the server racks.

To effectively eliminate these hotspots, DC operators sometimes overcool the rack environment [3]. However, over-cooling can lead to the creation of unexpected cold spots in the IT room surroundings. These unexpected hotspots and cold spots, also known as thermal anomalies can result in critical systems operating in unsafe temperature regions which in the long run increases server failure rate [4]. Moreover, the increase in server density of large HPC DCs can cause poor internal air flow and higher IT equipment temperatures leading to unstable reliability and potential increase in system downtime. Thereby, resulting in a significant rise in operating cost and Total Cost of Ownership (TCO) of DCs [5] [6].

Since temperature is considered one of the key indicators of the operational health of computing systems, the ability to accurately predict the thermal distribution of IT equipment is of utmost importance to DC operators. It provides them with the capability to ensure all systems are operating within the safe temperature range recommended by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) [7] [8].

So far, single-point temperature sensors such as thermocouples have been the popular choice for monitoring the operation states of servers. Despite their wide temperature range, their spatial resolution is insufficient in capturing detailed information regarding the thermal distribution of servers and racks. As a consequence, the use of infrared thermography (IRT) imager as thermal sensor for equipment condition monitoring has increased over the past few years. IRT imagers are capable of real-time monitoring of

the surface temperature distribution of equipment by detecting the radiated infrared energy from the surface. IRT imagers offer several benefits compared to single-point temperature sensors. These include but are not limited to their non-contact, anti-electromagnetic interference, high timeliness, non-destructive, and high accuracy characteristics. They are therefore, perfect candidates for the monitoring and diagnosis of server operation states.

By leveraging the benefits offered by IRT images, in this study, we evaluated seven state-of-the-art deep pretrained CNN-based algorithms and two shallow CNN-based algorithms for the automatic diagnosis of five server operation states. These operation states included CPU fan failure, main fan failure, server entrance blockage, partial CPU load, and maximum CPU load conditions. Our objective is to take advantage of the effective automatic feature extraction capabilities of CNNs and the rich temperature distribution information of the IRT images for the development of highly reliable and effective condition monitoring system of server operation states in DCs. This can reduce system down time, increase reliability, and reduce total cost of operation of DCs.

Specifically, this research paper makes the following main contributions:

- Integrates state-of-the-art CNN-based algorithms using transfer learning with infrared thermography for the improvement of prediction accuracy of server operation states including faulty conditions.
- Eliminates sources of human errors that emanate from handcrafted feature extraction techniques in previous studies by using more effective automatic CNN feature extraction techniques in the proposed models.
- Significantly increases server operation status diagnosis performance of the proposed CNN-based methods compared with baseline traditional machine learning approach with respect to F1-score, accuracy, recall, and precision metrics.

The results of this work can enable the proper workload planning and distribution among servers, increasing reliability and enhancing energy efficiency in DCs.

2 Related Work

Several studies have been conducted that attempt to offer solutions to the perplexing problem of thermal anomaly detection and thermal management in DCs.

The preponderance of these studies has focused on developing thermal management techniques that range from Computational Fluid Dynamics (CFD) modeling to data-driven modeling (DDM) and their hybrid counterparts to monitor and predict the thermal distribution of servers and hence decrease the high risk of failure of servers in DCs.

The CFD-based approaches use numerical algorithms to solve non-linear, three-dimensional differential equations that govern the thermal dynamics of

the DC internal environment. These CFD techniques are known to provide accurate results such as server inlet and outlet temperature prediction subject to various conditions [9]. However, the performance of CFD-based techniques relies on the consideration of several DC characteristics such as airflow rates, IT room dimensions, and setup [10]. In addition, CFD-based models are time-consuming and computationally expensive.

Recently, data-driven techniques have gained significant popularity as viable alternatives to CFD-based approaches for the modeling and prediction of system dynamics. This is triggered by the rapid increase in the availability of collected data of physical systems [11]. Data-driven approaches aim at finding relationships between the system state variables (input and output) without explicit knowledge of the physical behavior of the system [12] using multivariate statistical or machine learning-based techniques.

The benefits of using DDM techniques have been exploited by several researchers for the prediction of the temperature distribution of the IT room. One of such works is that of Athavale et al. [13]. They implemented and compared Artificial Neural Network (ANN), Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Proper Orthogonal Decomposition (POD) algorithms for the prediction of steady-state and transient-state rack inlet air temperature distribution in DCs using simulated data from a CFD model as input. Similarly, Ilager et al. [14] in their studies also explored several data-driven ML algorithms to accurately estimate the host temperature in a DC. Anastasiia et al., [15] implemented a K-means clustering ML algorithm for the localization of hotspots and the identification of distinct servers that regularly occurred in the overheated zones of the server room based on surrounding hot aisle ambient temperature data.

A few studies have attempted to combine CFD and DDM techniques in order to improve the prediction accuracy of temperature distribution in an IT room. One such study is that of Asgari et al., [16] whereby they combined ML algorithms with thermo-fluid transport equations to predict transient temperatures in server CPUs and server inlets.Their results showed that the performance of the hybrid model was superior to that of pure data-driven black-box and conventional zonal models. The reliability of DDMs on data obtained either from CFD simulations or from measurements using traditional temperature sensors such as thermocouples is one of the major drawbacks of such an approach. This is because single point temperature sensors can measure temperature only at a given point in space. Therefore, they are incapable of accurately capturing the spatiotemporal temperature changes inside a DC necessary for the successful analyses of hotspots. These can often result in high extrapolative error predictions [17].

To overcome the above limitation that stems from the use of data collected using single-point sensors or CFD models, and also improve the effectiveness of thermal management and accuracy of fault diagnosis of systems, some researchers have turned their attention to the use IRT. IRT offers a real-time,

non-contact and non-invasive approach to condition monitoring of civil structures, electrical installations, machineries and equipment [18]. This is due to its ability to monitor the temperature distribution of equipment surfaces by detecting the radiated infrared energy (heat) from the surface. And temperature is considered one of the most useful indicators of the structural health of equipment and components [19]. Some of the successful fault diagnosis applications using IRT include electrical equipment [20], rotating machinery [21], and induction motor inter turn [22].

In the domain of DCs, few studies have taken advantage of IRT by extracting handcrafted features from the IRT images and applying ML algorithms for system anomaly detection [6], temperature monitoring, and fault diagnosis. Liu et al. [23] [24] in their studies, extracted several features including texture features, Hu moments, morphological, statistical features, and modified entropy features from IRT images to diagnose various server operating states using an SVM classifier. Hu et al. [25] extracted corner and edge features from segmented IRT images. And by using an infrared fusion technique, their approach could identify the physical location of hotspots in a DC. Zhao et al. [26] developed a novel approach that extracts the correlations of sensing data from three sensors including a thermal camera, a microphone, and system performance logs for the prediction of hard disk drives failures of servers in DCs.

Each of the aforementioned studies employed a common two-step implementation process. First, they extracted features from the thermal images using handcrafted feature engineering methods. Second, they trained an ML algorithm(s) with the extracted features. This approach has some shortcomings. The prediction performances of DDM models are strongly correlated with the image preprocessing and hand-engineered feature extraction algorithms used. This acts as a bottleneck to the maximum achievable performance accuracies of the models.

Moreover, thermal images generally have a low signal-to-noise, low gray level, and low contrast compared with the normal visual RGB images [23]. Hence, a significant amount of time is usually spent in preprocessing and handcrafted feature extraction phases. Furthermore, extracting relevant features requires not only domain expertise but also several model training iterations with different combinations of feature sets to obtain high-performance accuracies.

The unprecedented rise in the development of state-of-the-art deep learning algorithms has offered researchers new techniques to overcome the limitations of handcrafted feature-trained ML algorithms and to boost performance. Specifically, CNNs, a class of deep learning, has proven highly effective in several computer vision tasks including image classification. This is due to the ability of CNNs to perform hierarchical learning, automatic feature extraction, multi-tasking, and weight sharing [27]. Despite the widespread application of CNN-based algoriths on IRT images for fault analysis is several fields, very

few studies have been conducted that implement a similar approach for automatic fault diagnosis in the DC infrastructure. Asgari et al. [28] combined a gray-box model with a CNN and RNN classifier for the diagnosis of cooling system failure in a DC. However, the source of input to their system model was not IRT images. It was data generated by the gray-box model.

Therefore, in this research study, we use thermal images captured with a FLIR E8 1.0 thermal imager as inputs to train and evaluate nine CNN-based models for the automatic detection, diagnosis, and classification of five server operation states: partial CPU load, maximum CPU load, main fan failure, CPU fan failure, and server entrance blockage.

To the best of our knowledge, this work is the first that comprehensively applies CNN-based algorithms to IRT images for the monitoring and diagnosis of the operation states of servers. It builds upon the work of Liu et al. [23] that applied handcrafted features to IRT images for the same purpose. It was further motivated by our preliminary studies [29] which illustrated the potential of deep neural networks and IRT for server fault classification.

The remainder of this research paper is organized as follows. In Section 3, we first give a research review of the CNN-based model architectures used in the study. Then, we dive deeper into the implementation methodology. In Section 4, we describe our performance evaluation approach, present the experimental results, compare results with those in the literature, and draw relevant insights for the data. Finally, Section 5 concludes the research study.

3 Methodology

In this section of the research study, we introduce the different CNN-based models that were explored in the detection, diagnosis, and classification of five server operation states using IRT images. These algorithms were grouped into two categories. Shallow Convolutional Neural Network models (S-CNN) and Deep convolutional neural networks (D-CNN). The first S-CNN model consists of a single convolution layer (1-CNN model) classifier. The second S-CNN model consists of two convolutional layers (2-CNN). On the other hand, the D-CNN models included seven pre-trained deep convolutional models: ResNet50, ResNet34, VGG-19, MobileNetV2, DenseNet121, DenseNet201, and AlexNet. Before introducing various CNN-based models implemented in this work, we first give a brief and succinct overview of the building blocks that are the foundations of the deep CNN-based models.

3.1 Convolutional Neural Network Architecture

CNN is a class of Artificial Neural Networks (ANN) that saw the spotlight through the groundbreaking work of Yann Lecun in 1989 [30]. Their powerful ability to efficiently process and learn highly representative, hierarchical features from a diverse set of training data such as images and time-series data has drawn an unprecedented amount of attention to them in recent years.

The architecture of a CNN is made up of multiple layers: convolutional layers, non-linearity layers, pooling layers, and fully-connected layers with specific functions (cf. Fig. 1). The convolutional layer and the fully-connected layers have trainable parameters. However, the pooling (subsampling) layer and non-linearity do not have parameters.

CNN-based models offer certain unique advantages compared with traditional ML algorithms. First, in most applications of CNNs that use images as inputs, relatively little preprocessing of data is required prior to training. Second, their ability to automatically extract useful low and high level features prevents them from the drawbacks associated with the design of handcrafted features that is subjective to user experience. Next, we give a brief description of the main layers that make up a typical CNN architecture.

3.1.1 Convolutional Layer

he convolutional layer is composed of a set of convolutional kernels (filters) where each neuron acts as a kernel [31]. Convolutional kernels work by dividing the image into small slices, commonly known as receptive fields. Mathematically, the convolution operation is expressed in (1) as follows:

$$F_l^k(p,q) = \sum_n \sum_{x,y} I(x,y) .a_l^k(u,v) \tag{1}$$

Where I(x,y) represents a unit of the input image tensor I and a_l^k(u,v) is the index of the *k-th* convolutional kernel. The operation between both terms is element-wise multiplication.

3.1.2 Pooling Layer

The pooling layer also referred to as the down-sampling layer typically follows the convolutional layer in a traditional CNN architecture. This layer is used to reduce the dimension of the previous feature map while ensuring the extracted features are invariant to translational shifts and small distortions. The pooling layer aggregates information in the small region of the receptive field of the input feature channels and outputs the dominant response within this small region [32]. Each pooled feature map corresponds to one feature map of the previous layer. Some existing pooling operation techniques include Max-Pooling, Global Pooling, Global Max-Pooling, Average Pooling, etc. Assuming that F_l^k is the input feature map to the pooling layer, the pooling operation is expressed in (2) as follows:

$$Z_l^k = g_p\left(F_l^k\right) \tag{2}$$

F_l^k is the k^{th} feature map of the l^{th}input layer. Z_l^k is the extracted pooled feature map of the input feature map F_l^k. And, $g_p(F_l^k)$ represents the particular pooling operation used.

3.1.3 Fully-Connected Layer

The fully-connected layer (FC) takes as input a one-dimensional feature vector that is a concatenation of the feature maps of images obtained from the previous sequential convolutional and pooling operations. The FC layer is often located at the end of the CNN architecture where it connects every neuron in one layer to every neuron in the next layer. The output of the fully connected layer can be obtained by making a weighted summation of the input and response by the activation function [33], as stated mathematically in equation (3)

$$X_l = f\left(W_l * F_l^k + b_l\right) \tag{3}$$

X_l is the output obtained by operations of a non-linear activation function $f(x)$. For most CNN-based architectures such as the ones used in this study, $f(x)$ is typically a rectified linear unit (ReLU) or a sigmoid activation function. W_l is the l layer feature map corresponding weight matrix. F_l^k denotes the l^{th} layer's k^{th} feature map and b_l is the bias term.

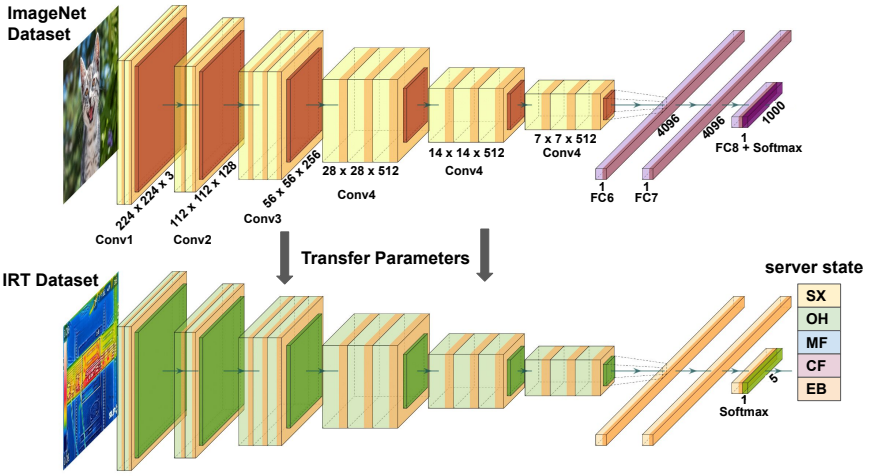

Fig. 1 Transfer learning process for server operation state classification. The top model architecture is a VGG-19 base model with convolutional, ReLU and pooling layers, that is pretrained on the ImageNet dataset. Features extracted from the base model are transfered to the target CNN model (bottom) to be trained on the IRT images for the diagnosis of each of the five server states.

3.2 Deep CNN-based Model Architectures

In this section, we describe in brief the architecture of the seven D-CNN models implemented in this work for the diagnosis of the five server operation states using IRT images.

3.2.1 Residual Neural Network Model

Residual neural network (ResNet) is a groundbreaking CNN architecture in the domain of deep learning proposed by He et al. [34]. The key attribute of ResNets is that they can solve the notorious vanishing gradient problem that plagued previous deep CNN model architectures. The ResNet network architecture is based on the deep residual framework, which uses shortcut connections (skip connections). Several variants of the ResNet architecture exist with the main difference being the number of layers. In this research study, we explored two ResNet model architectures: ResNet-50 and ResNet-34 pre-trained models. The ResNet-50 model architecture consists of five stages each with a convolution and Identity block. The convolution and the Identity block are both made up of three convolutional layers. The Resnet-50 and ResNet-34 models adopted for this study were pre-trained with the ImageNet dataset [35] and then fine-tuned and adapted for the diagnosis of the five server operation states using IRT images.

3.2.2 Visual Geometry Group Model

The Visual Geometry Group (VGG) is a CNN model architecture introduced by Zisserman et al. [36]. There exist several variants of the VGG architecture, such as VGG-11, VGG-16, VGG-19, etc. In this study, we adopted the VGG-19 model architecture that consists of 19 layers (16 convolution layers, three fully connected layers, five MaxPool layers and one SoftMax layer). Specifically, the VGG-19 model imported using the Fastai library [37] was pre-trained using images from the ImageNet dataset and then used during the training of the IRT images.

3.2.3 Densely Connected Convolutional Networks Model

Densely Connected Convolutional Networks (DenseNet) is a classical CNN model proposed by Huang et al. in 2017 [38]. DenseNet comprises mainly of Dense Blocks and transition layers. The Dense Block allows the concatenation of the output of the previous layer with the future layer. The transition layer regulates the number of channels ensuring it is kept relatively small. There exist different variants of the DenseNet architecture including DenseNet-121, DenseNet-160, DenseNet-201, etc. The trailing numbers represent the number of layers in each version of DenseNet. In this study, we implemented DenseNet-121 and DenseNet-201 networks with pre-trained weight obtained from training the models with the ImageNet dataset.

3.2.4 Google MobileNet Model

The MobileNet model architecture was developed by Andrew G. Howard et al. [39]. The model aimed at effectively maximizing the accuracy while taking into consideration the restricted resources for an on-device or embedded application. MobileNets are designed to meet resource constraints for a variety of

use cases. This is because they are small, low-latency, low-power models. Currently, there exist three versions of the MobileNet architecture. In this work, we implemented MobileNet version 2 (MobileNet-V2) with pre-trained weights on the ImageNet database. MobileNet-V2 consists of two main blocks. Residual block with a stride of one and a second block with a stride of two for downsampling.

3.2.5 AlexNet Model

The AlexNet model, considered as one of the most influential models in the field of computer vision, was introduced by Alex Krizhevsky et al. [40] in 2012. AlexNet model architecture consists of five convolutional layers with max-pooling layers sandwiched between some of them. The convolutional layers are then followed by three fully-connected layers. Compared with today's deeper models, it is a relatively shallow CNN model. Although, in this study, AlexNet falls among the deep CNN model class as it is relatively deeper compared to the shallow models. We deployed the AlexNet model initialized with pre-trained weights using the ImageNet database.

4 Proposed Methodology

This section presents, in detail, the step-by-step approach that was followed for the successful diagnosis of the five server operation states in a DC.

The fundamental building block of both the S-CNN and D-CNN models is the convolutional layer. See section 3.1 above for more details. By stacking these convolutional layers with the other layers, both S-CNN and D-CNN models are capable of effectively extracting low, mid, and high-level features from the IRT images to accurately and automatically diagnose the different server operation states.

The core steps that formed the backbone of our implementation strategy for the detection and classification of the different server states using IRT images are: 1. Infrared thermal image dataset acquisition, 2. Infrared thermal image preprocessing, 3. Models training and validation with stratified cross-validation, 4. Performance evaluation of models for fault detection.

4.0.1 Infrared Thermal Image Dataset Acquisition

The dataset used in this research was collected using a FLIR E8 1.0 IRT imager by Lui et al. [23]. The dataset is not publicly available. It was privately shared with the authors of this research work upon request. For the complete experimental setup, data collection procedure and IRT camera specifications, refer to [23], and [24] for more details.

The dataset contained a total of 1350 thermal image samples divided into five classes. Each class represented a specific state of operation of the server. The five states are partial CPU load (SX - 60% CPU load), maximum CPU load (OH - 100% CPU load), main fan failure (MF), CPU fan failure (CF), and

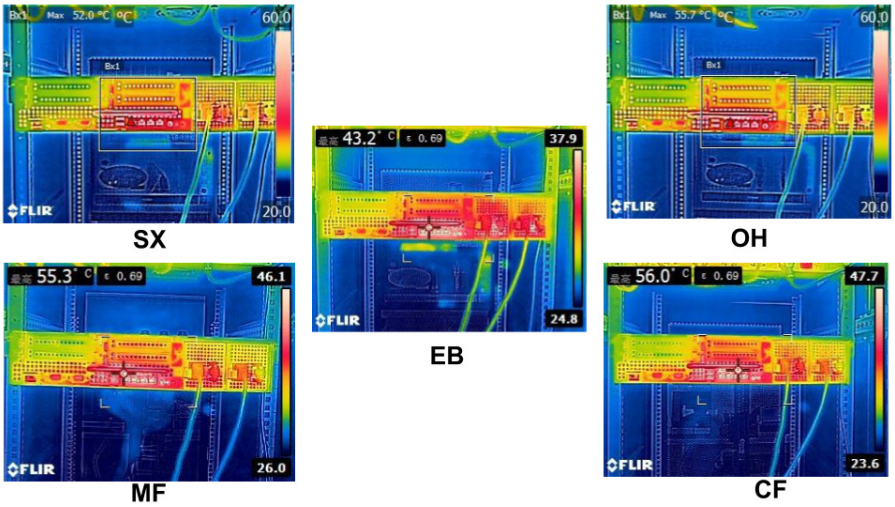

Fig. 2 Sample of captured IRT thermal images for each of the five server operation states (MF: main fan failure, SX: partial CPU load, OH: maximum CPU load, EB: entrance blockage, CF: CPU fan failure)

server entrance blockage (EB). Each class have a total of 270 image sample. Figure 2 shows one IRT image samples drawn from each class.

4.0.2 Infrared Thermal Image Dataset Preprocessing

As mentioned above, one of the main advantages of CNN-based models is the minimal dataset preprocessing they require prior to training. In this implementation, we applied image resizing, data augmentation and data normalization preprocessing steps to the IRT images.

Image Resizing : Resizing is a critical preprocessing step in the building of deep CNN models. The raw thermal images contained in the dataset were RGB images of original dimensions 240×320. Each image was resized to a new square dimension of 224×224. By resizing the images the models trained more efficiently. Also, the pre-trained deep CNN models expect inputs of dimension 224×224 as they were originally constructed and trained using images from the ImageNet database of the same dimension. After resizing, we passed the smaller dimension images down the pipeline for further preprocessing.

Data Augmentation: Given the relatively small size of our dataset (1350 images), we needed to augment our dataset to avoid overfitting our models. We applied geometric transformations such as horizontal flip, rotations, zooming, warping, etc., on the resized thermal images.

Data Augmentation: The last preprocessing technique we applied to the dataset prior to training the models was the data augmentation. The goal of normalization was to transform features to be on a similar scale. This helped to improve model performance and training stability. By importing the image statistics (mean and standard deviation) of the images from the ImageNet

database through the FastAI library, the IRT images were normalized. Thus, completing the preprocessing phase.

4.0.3 Model Training with Stratified Cross-Validation.

In this phase of the implementation pipeline, we trained both the S-CNN models and D-CNN models independently using a technique known as stratified cross-validation (CV). Stratified CV is a variant of cross-validation. Cross-validation is a data resampling technique that enables the objective assessment of the generalization ability of predictive models and also prevents overfitting [41]. Furthermore, it leads to a better average performance at the same time protects against the possibility of disastrous performance [42].

In stratified CV, the dataset splitting into folds is based on a criterion, which ensures that each fold contains the same number of observations with respective categorical values such as the class outcome value. K is a hyper-parameter that is chosen during training and represents the number of folds for which the dataset is split. In this study, we chose a K value of five. Hence, implementing a stratified 5-fold CV.

The IRT image dataset constituted a total of 1350 thermal images divided into 5 classes of 270 samples each. Each class represented a specific server operating condition (cf. Fig. 2). For an objective performance evaluation of the various model classifiers, the IRT dataset was divided into five different data split ratios (50/50, 60/40, 70/30, 80/20, 90/10) used for training and testing. Stratified 5-fold CV was applied during training and the average training and validation performance measured. The test split was used to estimate the unbiased diagnosis performance of the trained models.

The models were run on a Google Colab notebook that host a free NVIDIA Tesla K80 Graphics Processing Unit (GPU) with a 12GB RAM provided by Google cloud services. The programming language used for this study is the Python programming language.

Training Shallow-CNN Models

The two shallow CNN algorithms (single-layer CNN and two-layer CNN) were built and trained using stratified 5-fold CV. These models were implemented using the Keras open-source software library. Each model was trained for 10-epochs with a batch size of 32.

The objective of training these shallow CNN models is to give a complete comparative analysis of their performance with respect to the more advanced DL models and traditional ML algorithms. Also, we aimed at exploring the advantages these simple models could offer in the diagnosis, detection, and classification of the thermal state of servers in DCs in terms of efficiency and accuracy.

Training Deep Pre-trained CNN Models - Transfer Learning

The deep CNN-based architectures applied in this study for server fault diagnosis consists of many feature-learning layers. Training such architectures from scratch is a challenging task. This is because they have large number of weights which are randomly initialized before the training process and iteratively updated based on labeled data and the loss function [43]. The weight update process is therefore time-consuming. This situation is exarcebated by the limited number of IRT image samples, making these models susceptible to overfitting.

To overcome these drawbacks, we applied the technique of transfer learning by using pretrained deep CNNs, that have been pretrained on the ImageNet dataset which consists of more than a million samples of natural images.

Transfer learning aims to improve the learning of a target task through the transfer and utilization of previously-acquired knowledge from a related source task [44][45]. Based on the notations used in [45], given a source domain $\mathcal{D}_S = \{\mathcal{X}_S, P(X_S)\}$, a source task $\mathcal{T}_S = \{Y_S, f_S(.)\}$, a target domain $\mathcal{D}_T = \{\mathcal{X}_T, P(X_T)\}$, and a target task $\mathcal{T}_T = \{Y_T, f_T(.)\}$, transfer learning aims at improving the learning performance of $\mathcal{T}_T$ when $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. Where $\mathcal{X}_S$ and $\mathcal{X}_T$ represent the source and target data samples. $P(X_S)$ and $P(X_T)$ are their corresponding marginal probability distributions. Y_S and Y_T represents the source and target label space. $f_S(.)$ and $f_T(.)$ are the source and target predictive functions.

In the particular case of server state diagnosis, $\mathcal{X}_S$ represents labeled images from the large ImageNet database and $\mathcal{X}_T$ consist of the labeled IRT images. Transfer learning helps to speedup the training procedure, improve learning of hierarchical features, and boost the performance of deep CNNs.

In our approach, transfer learning was applied by removing the last layers of each pretrained D-CNN model that was customized for the ImageNet dataset classification task, and replacing it with new layers fit for the classification of the five server operation states. First, the weight parameters of the lower level convolutional layers of the pretrained models were frozen. These layers capture lower-level features such as edges and gradients common to most image classification tasks making them useful. Second, we updated the weight parameters of the new added layers using the IRT image samples through training. This approach allowed the knowledge learned from the millions of datapoints of the ImageNet database to be transfered to the task of server fault diagnosis in DCs (cf. Fig. 1). Using pre-trained weights also accelerated the training process of the models on the new task. Furthermore, it improved the accuracy of the classifiers. Each D-CNN model was trained for 5 epochs with a specific learning rate derived from experimentation to speed up training and model convergence.

4.0.4 Performance Evaluation.

In this study, standard ML evaluation metrics were used for the evaluation of the performance of both the D-CNNs and S-CNNs on the diagnosis of the five server states. These metrics include precision, recall, F1-score, and accuracy. They are mathematically expressed in equations (4) - (7).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

where TP, TN, FP, FN correspond to True Positives, True Negatives, False Positives, False Negatives, respectively.

5 EXPERIMENTAL RESULTS AND DISCUSSIONS

The monitoring of the thermal operation states of servers requires highly reliable, effective, and efficient monitoring systems. By implementing seven D-CNN models with transfer learning and two S-CNN models using IRT images, we aimed at developing a diagnostic system that meets such requirements.

5.1 Diagnosis of Server Operation States with varying Dataset Sizes

In principle, the performance of deep learning models are influenced by the size of the training dataset. Specifically, research shows that the larger the number of training samples, the better the prediction ability of CNN-based models. Therefore, to evaluate the performance of the nine CNN-based models, we performed a comparison analysis of the models based on the varying size of the training IRT images.

The validation accuracy and corresponding standard deviation was used as a metric to both evaluate the fault diagnostic performance of each model and also perform a comparison analysis of the models across the five IRT dataset split ratios. Figures 3 and 4 show the training and validation accuracies with the corresponding standard deviation values of each CNN-based model.

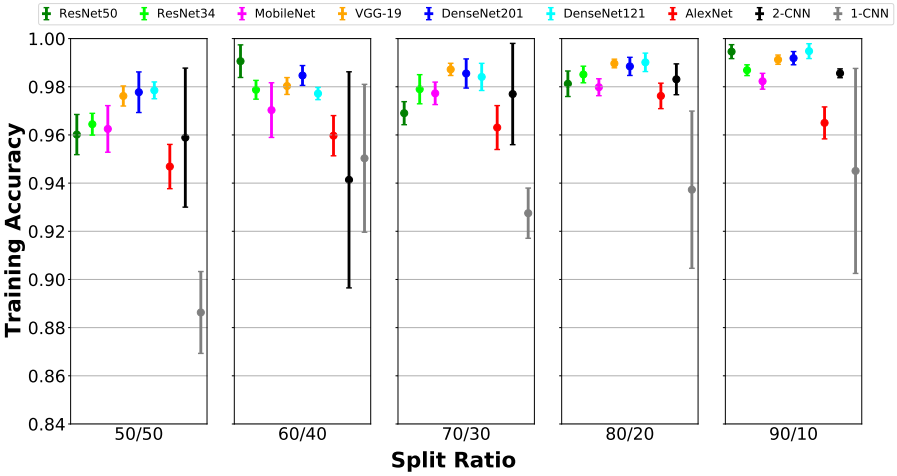

Fig. 3 Training set accuracies and respective standard deviations for five split ratios on IRT images. D-CNNs have higher training accuracies compared with S-CNNs. The S-CNNs have higher standard deviation values compared with D-CNNs as indicated by the relatively longer bars. Hence, S-CNN are more sensitive to overfitting.

Based on our experimental results, DenseNet121 achieves the highest average validation accuracy (98.92%) with an average validation standard-deviation of 0.0067 across the five dataset sizes. Therefore, it is considered the overall best CNN-based model classifier for diagnosis of server operation states using IRT images. AlexNet obtained the least average validation accuracy of 97.63% with a corresponding average standard deviation of 0.0094 among the seven pre-trained deep CNN-based models. The 1-CNN model classifier obtained the overall least average validation accuracy score of 92.43% with a corresponding highest average standard deviation of 0.0288. Hence, it is the worse model classifier among the implemented nine CNN-based models.

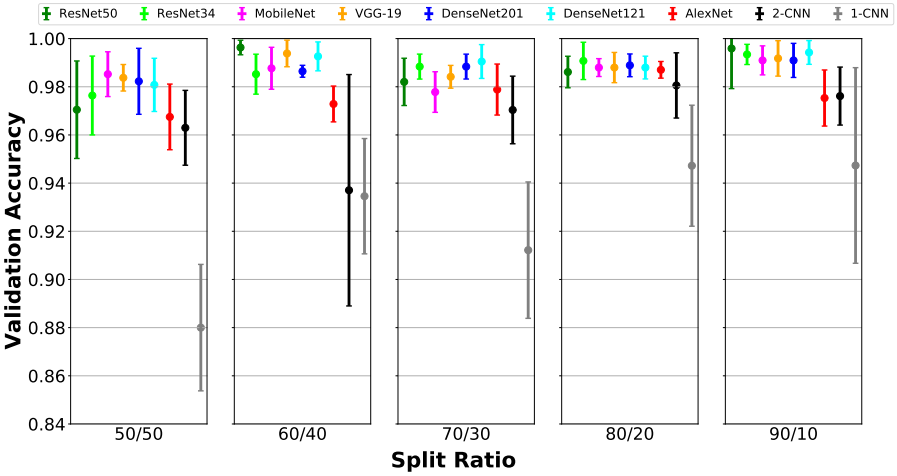

Fig. 4 Validation set accuracies and respective standard deviations for five split ratios on IRT images. D-CNNs have higher training accuracies compared with S-CNNs. However, they have lower standard deviation values compared with D-CNNs as indicated by their relatively shorter bars. Hence, better model classifiers server operation state prediction.

These results highlight the superiority and advantage of using deep pretrained CNN-based architectures compared to shallow CNN-based architectures especially in the case of limited availability of training samples. Furthermore, pretrained CNN-based models achieved remarkable performance with a relatively small number of training epochs and lesser training times.

To obtain an unbiased estimation of the performance of the CNN-based model classifier for the five split ratios, we used the holdout test set. The unseen test set was fitted to the final trained models and the confusion matrix was used to evaluate their prediction performance. The confusion matrix assesses the performance of each model by comparing the true server operation state with the predicted state. Based on the confusion matrix, the test accuracy, precision, recall, and F1-score were computed.

In Figures 5, 6, 7, and 8, which respectively represent the test accuracy, precision, recall, and F1-score, computed from the confusion matrix, we observe that all model classifiers show good prediction capabilities of server operation states across the five split ratios. In the case of the 90/10 split ratio, DenseNet201 and VGG-19 correctly predicted all five server operation states achieving a 100% test accuracy.

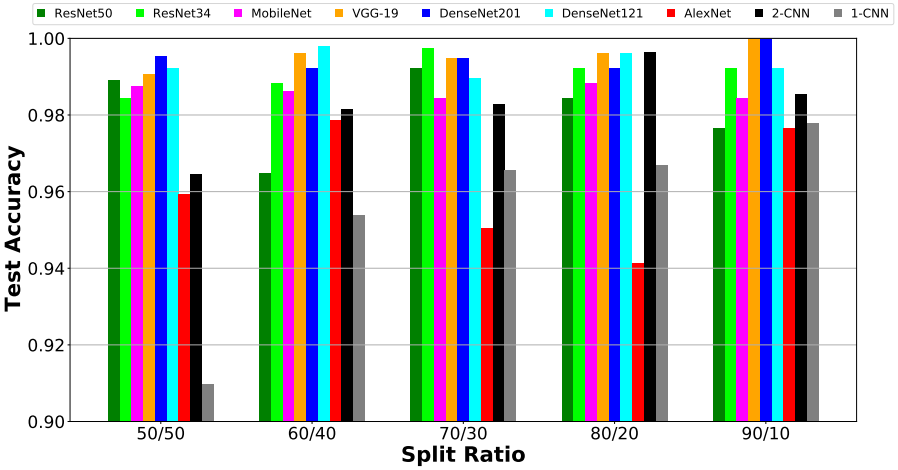

Fig. 5 Test accuracy of CNN-based model classifiers on holdout test set for five split ratios. Among the D-CNNs, AlexNet has the lowest test accuracy score for all split ratios except for the 60/40 split ratio. Among the S-CNNs, the 2-CNN model classifier has higher test accurary compared with 1-CNN model classifier for all five split ratios.

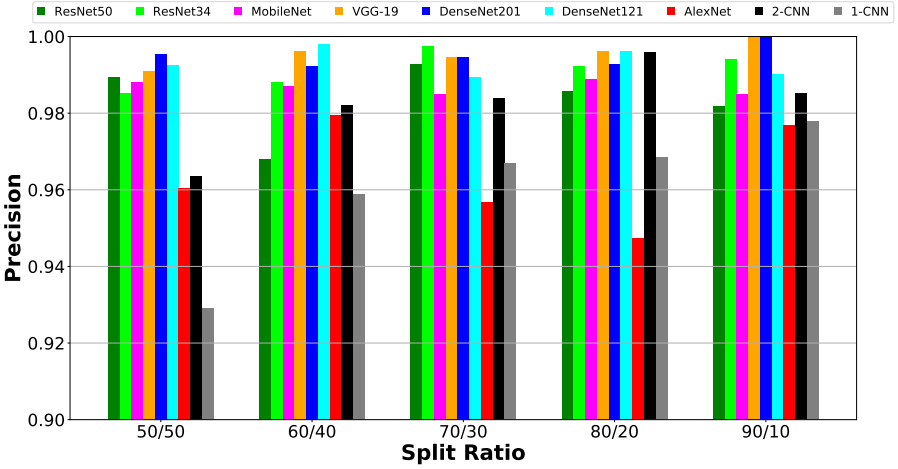

Fig. 6 Precision scores of CNN-based model classifiers on holdout test set for five split ratios. Depending on the split ratio, some model classifiers have higher precision scores than others. However, among the D-CNNs, AlexNet has the lowest precision score for all split ratios except for the 60/40 split ratio. Among the S-CNNs, the 2-CNN model classifier has higher precision score compared with 1-CNN model classifier for all five split ratios.

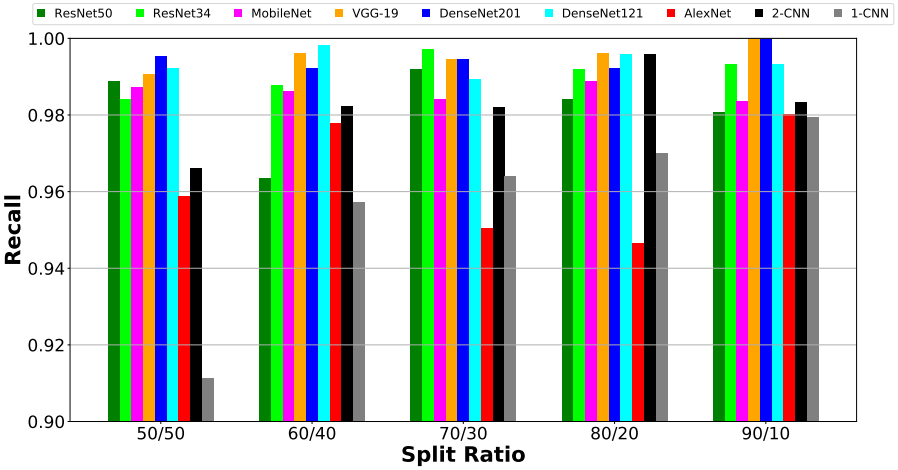

Fig. 7 Recall scores of CNN-based model classifiers on holdout test set for five split ratios. Some model classifiers have higher precision scores than others based on split ratio. But, among the D-CNNs, AlexNet has the lowest precision score for all split ratios except for the 60/40 split ratio. Among the S-CNNs, the 2-CNN model classifier has higher precision score compared with 1-CNN model classifier for all five split ratios.

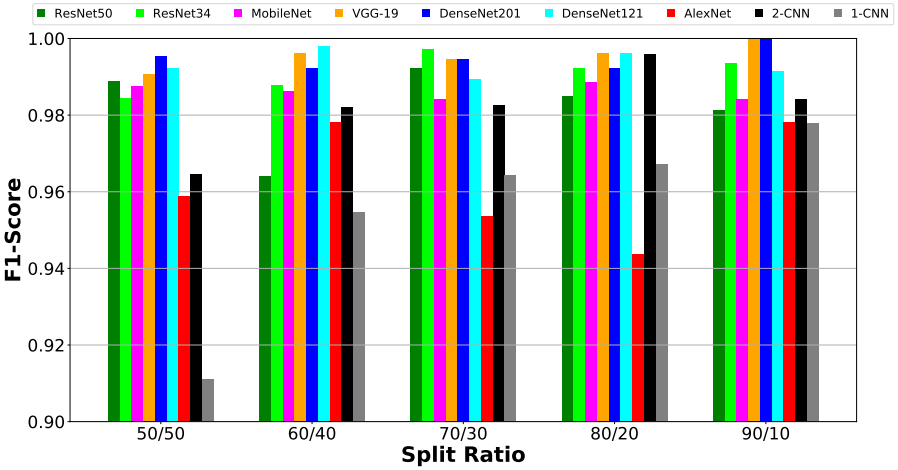

Fig. 8 F1-scores of CNN-based model classifiers on holdout test set for five split ratios. The D-CNNs have higher F1-score values compared with the S-CNNs for all five splits with the exception of AlexNet in the 70/30 and 80/20 split ratio. Overall, D-CNNs are better at server operation state diagnosis.

Based on the average test accuracy, precision, recall, and F1-score across all five splits, VGG-19 model outperformed all models for the server operation state prediction on the unseen IRT images. Among the D-CNN models, AlexNet had an overall lowest test performance score for all split ratios except

for the 60/40 split ratio. The shallow 2-CNN model classifier's prediction performance on the test set for a split ratio of 80/20 was remarkably high and matched those of advanced D-CNN model classifiers despite the fact that it was not pretrained on the ImageNet dataset and has only 2 convolutional layers in its architecture. Finally, the 1-CNN was the worst model in the prediction of server operation states in data centers. This is because 1-CNN model classifier with only 1 convolutional layer cannot capture sufficient complex and meaningful features from the limited IRT images, thereby, resulting in more server operation state misclassifications.

The 80/20 split ratio is considered to produce the overall best validation and test performance among all split ratios. This can be seen by the number of model classifiers whose performance test accuracy, precision, recall, and F1-score values lie above the 98% horizontal line (cf. Figures 5 - 8). The 70/30 split ratio generated almost similar test performance results as the 80/20 split ratio. This is in line with results obtained by Nguyen et al. [46], who investigated the influence of data splitting on the performance of three ML models in the prediction of shear strength of the soil. They found that 70/30 was the best split ratio.

The robustness of the prediction ability of the CNN-based models with varying training IRT dataset sizes was measured based on the average standard deviation values of the validation accuracy for each model across the five splits. From Figure 9, we can see that VGG-19 had the lowest average standard deviation of 0.0059 and hence was the most robust in terms of performance degradation with varying size of training set. 1-CNN model had the highest average standard deviation of 0.0288. These results indicate that D-CNN model classifiers are robust with higher performance accuracies and lower variability, hence, most reliable for automatic server operation state diagnosis. The S-CNN model classifiers, on the other hand, are more sensitive to changes in the training size of the IRT images and also they are susceptible to overfitting.

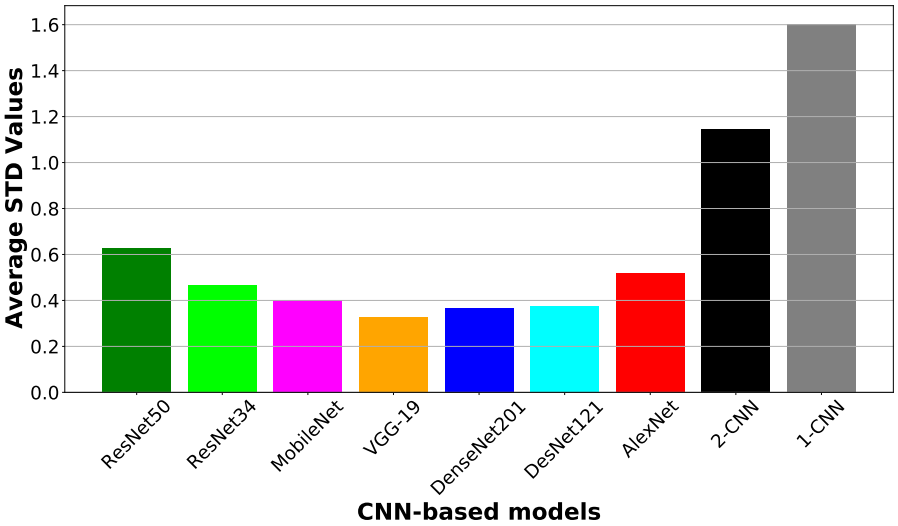

Fig. 9 Average validation set standard deviation (STD) across five split ratios.The lower the value the better the model. Shallow CNN-based models have the highest validation set standard deviation values, hence they are more sensitive to overfitting. VGG-19 has the lowest validation set standard deviation value and hence is considered the most robust model.

5.2 Comparison of CNN-based models with traditional ML models.

The objective of this section is to compare the performance of our deep learning CNN-based approach to the machine learning approach used in the literature for the automatic diagnosis of the five types of server operating conditions with the same IRT dataset. Specifically, in the work of Lui et al. [23], they extracted handcrafted features such as morphological features, texture features and statistical features from the IRT images and then trained a support vector machine (SVM) algorithm to diagnose the final status of the server. In their experiment, the IRT dataset was split into a training set of 900 IRT images (67%) and a test set of 450 IRT images (37%).

We used the test accuracy as the metric of performance comparison as that is the metric used in the literature. To ensure no unneccessary bias in performance of our models based on number of training and testing samples, we used the models trained with the 60/40 split ratio. That is, training set of 810 IRT images and test set of 540 IRT images. This means the CNN-based models were trained with slightly lesser number of IRT images compared to that of Liu et al [23]. However, as shown in Table 1, the CNN-based models outperformed the SVM classifier that had a test accuracy of 91.11% [23]. The DenseNet121 model classifier had the highest test accuracy of 99.80%. This corresponds to a percentage increase of approximately 9.54% compared to the SVM classifier. Among all CNN-based model classifiers, 1-CNN has the lowest

test accuracy of 95.38%. However, it outperformed the SVM classifier trained on hand-crafted features by a margin of approximately 4 percentage points.

This significant performance improvement in our CNN-based approach compared with traditional ML approach for the diagnosis of server faults in DCs is due to the powerful ability of CNN-based models to automatically learn low and high thermal features of the temperature distribution of the server surface for each operation state.

Models	ResNet50	ResNet34	MobileNet	VGG-19	DenseNet201	DenseNet121	AlexNet	2-CNN	1-CNN	SVM
Test Acc(%)	96.48	98.83	98.63	99.61	99.22	99.80	97.85	98.15	95.38	91.11

Table 1 Test set accuracy for 60/40 split ratio on IRT images. Higher is better. All CNN-based models have higher test set accuracy compared with the traditional SVM classifier used in [23]. DenseNet121 model outperforms the rest of the models

6 CONCLUSION

In this study, we investigated two shallow and seven state-of-the-art deep pretrained CNN-based architectures for the monitoring and diagnosis of five server operation conditions including two faulty conditions, based on the thermal distribution information of the server surface captured using an IRT imager. A transfer learning approach was adopted for the deep CNN model classifiers. Each model's predictive performance was evaluated using a stratified 5-fold cross-validation resampling technique across five dataset split ratios.

The deep pretrained CNN-based models achieved higher diagnosis accuracies of the server operation condition compared with shallow CNN-based models. In particular, the pretrained DenseNet121 architecture had the highest average validation accuracy of 98.83% with an average standard deviation value of 0.0067 across all five IRT image dataset split ratios. Therefore, it was considered the most effective and robust model for the task of fault diagnosis of servers in DCs.

In addition, a performance comparison of the nine CNN-based architectures with an SVM classifier trained using handcrafted features showed higher test prediction accuracies of the CNN-based model classifiers with an approximate maximum and minimum percentage point margin of 8 and 4.

These experimental results indicate that deep CNN-based model architectures applied on IRT images can provide DC operators with an effective non-contact intelligent tool for server operation state monitoring and diagnosis. Furthermore, by using the proposed approached, DC operators can monitor the server operation states with higher accuracy and improve the thermal management, thereby increasing the reliability and energy efficiency of the DC infrastructure.

Acknowledgments. We are grateful to Hang Liu for providing us the infrared thermal image dataset used for the successful implementation of this research study upon request. This work would not have been possible without

the dataset. This work is supported by Istanbul Technical University (ITU) Vodafone Future Lab under Project Number 2018000463.

References

[1] Lin, M., Shao, S., Zhang, X.S., VanGilder, J.W., Avelar, V., Hu, X.: Strategies for data center temperature control during a cooling system outage. Energy and Buildings **73**, 146–152 (2014)

[2] Ellsworth Jr, M.J., Singh, P., Chu, R.C., *et al.*: Liquid cooling architectures for computer systems of high availability. ASHRAE Transactions **113**, 136 (2007)

[3] Khalaj, A.H., Halgamuge, S.K.: A review on efficient thermal management of air-and liquid-cooled data centers: From chip to the cooling system. Applied energy **205**, 1165–1188 (2017)

[4] Srinivasan, J., Adve, S.V., Bose, P., Rivers, J.A.: The impact of technology scaling on lifetime reliability. In: International Conference on Dependable Systems and Networks, 2004, pp. 177–186 (2004). IEEE

[5] Fakhim, B., Behnia, M., Armfield, S., Srinarayana, N.: Cooling solutions in an operational data centre: A case study. Applied thermal engineering **31**(14-15), 2279–2291 (2011)

[6] Lee, E.K., Viswanathan, H., Pompili, D.: Model-based thermal anomaly detection in cloud datacenters using thermal imaging. IEEE Transactions on Cloud Computing **6**(2), 330–343 (2015)

[7] ASHRAE: Technical committee 9.9, thermal guidelines for data processing environments. American Society of Heating, Refrigerating and Air-Conditioning Engineers, **4th ed. Atlanta: W. Stephen Comstock** (2015)

[8] Lin, P.: How to fix hot spots in the data center. https://download.schneider-electric.com/files?pDocRef=SPDVAVR-9GNNGREN. [Online; accessed 19-June-2021]

[9] Choi, J., Kim, Y., Sivasubramaniam, A., Srebric, J., Wang, Q., Lee, J.: A cfd-based tool for studying temperature in rack-mounted servers. IEEE transactions on computers **57**(8), 1129–1142 (2008)

[10] Zapater, M., Risco-Martín, J.L., Arroba, P., Ayala, J.L., Moya, J.M., Hermida, R.: Runtime data center temperature prediction using grammatical evolution techniques. Applied Soft Computing **49**, 94–107 (2016)

[11] Montáns, F.J., Chinesta, F., Gómez-Bombarelli, R., Kutz, J.N.: Data-driven modeling and learning in science and engineering. Comptes Rendus Mécanique **347**(11), 845–855 (2019)

[12] Solomatine, D.P., Ostfeld, A.: Data-driven modelling: some past experiences and new approaches. Journal of hydroinformatics **10**(1), 3–22 (2008)

[13] Athavale, J., Yoda, M., Joshi, Y.: Comparison of data driven modeling approaches for temperature prediction in data centers. International Journal of Heat and Mass Transfer **135**, 1039–1052 (2019)

[14] Ilager, S., Ramamohanarao, K., Buyya, R.: Thermal prediction for efficient energy management of clouds using machine learning. IEEE Transactions on Parallel and Distributed Systems **32**(5), 1044–1056 (2020)

[15] Grishina, A., Chinnici, M., Kor, A.-L., Rondeau, E., Georges, J.-P.: A machine learning solution for data center thermal characteristics analysis. Energies **13**(17), 4378 (2020)

[16] Asgari, S., MirhoseiniNejad, S., Moazamigoodarzi, H., Gupta, R., Zheng, R., Puri, I.K.: A gray-box model for real-time transient temperature predictions in data centers. Applied Thermal Engineering **185**, 116319 (2021)

[17] Asgari, S., Moazamigoodarzi, H., Tsai, P.J., Pal, S., Zheng, R., Badawy, G., Puri, I.K.: Hybrid surrogate model for online temperature and pressure predictions in data centers. Future Generation Computer Systems **114**, 531–547 (2021)

[18] Bagavathiappan, S., Lahiri, B., Saravanan, T., Philip, J., Jayakumar, T.: Infrared thermography for condition monitoring–a review. Infrared Physics & Technology **60**, 35–55 (2013)

[19] Epperly, R.A., Heberlein, G.E., Eads, L.G.: A tool for reliability and safety: predict and prevent equipment failures with thermography. In: Record of Conference Papers. IEEE Industry Applications Society 44th Annual Petroleum and Chemical Industry Conference, pp. 59–68 (1997). IEEE

[20] Jadin, M.S., Taib, S.: Recent progress in diagnosing the reliability of electrical equipment by using infrared thermography. Infrared Physics & Technology **55**(4), 236–245 (2012)

[21] Janssens, O., Schulz, R., Slavkovikj, V., Stockman, K., Loccufier, M., Van de Walle, R., Van Hoecke, S.: Thermal image based fault diagnosis

for rotating machinery. Infrared Physics & Technology **73**, 78–87 (2015)

[22] Singh, G., Kumar, T.C.A., Naikan, V.: Induction motor inter turn fault detection using infrared thermographic analysis. Infrared Physics & Technology **77**, 277–282 (2016)

[23] Liu, H., Bao, C., Xie, T., Gao, S., Song, X., Wang, W.: Research on the intelligent diagnosis method of the server based on thermal image technology. Infrared Physics & Technology **96**, 390–396 (2019)

[24] Liu, H., Xie, T., Ran, J., Gao, S.: An efficient algorithm for server thermal fault diagnosis based on infrared image. In: Journal of Physics: Conference Series, vol. 910, p. 012031 (2017). IOP Publishing

[25] Hu, J.-J., Li, H.-C., Tai, H.-M.: Thermal distribution monitoring of the container data center by a fast infrared image fusion technique. Computers & Mathematics with Applications **64**(5), 1484–1494 (2012)

[26] Zhao, M., Furuhata, R., Agung, M., Takizawa, H., Soma, T.: Failure prediction in datacenters using unsupervised multimodal anomaly detection. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 3545–3549 (2020). IEEE

[27] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: A review. Neurocomputing **187**, 27–48 (2016)

[28] Asgari, S., Gupta, R., Puri, I.K., Zheng, R.: A data-driven approach to simultaneous fault detection and diagnosis in data centers. Applied Soft Computing **110**, 107638 (2021)

[29] Beltus, N.W., Hamza, S.E., Behcet, U.T.: A deep learning approach to fault detection and classification in datacenters. In: Basarim High Performance Conference, 2020 (2020)

[30] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)

[31] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, ??? (2016)

[32] Lee, C.-Y., Gallagher, P.W., Tu, Z.: Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: Artificial Intelligence and Statistics, pp. 464–472 (2016). PMLR

[33] Bouvrie, J.: Notes on convolutional neural networks (2006)

[34] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image

recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[35] ImageNet Database. https://www.image-net.org/. [n.d. Available Online]

[36] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[37] Jeremy, H., Sylvain, G.: Fastai Python Library Version 1.0.61. http://docs.fast.ai/. [Software Library]. [Online; accessed 5-May-2021] (2020)

[38] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[39] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[40] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)

[41] Ranganathan, S., Nakai, K., Schonbach, C.: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Elsevier, ??? (2018)

[42] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 1–48 (2019)

[43] Shao, S., McAleer, S., Yan, R., Baldi, P.: Highly accurate machine fault diagnosis using deep transfer learning. IEEE Transactions on Industrial Informatics **15**(4), 2446–2455 (2019). https://doi.org/10.1109/TII.2018.2864759

[44] Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 242–264. IGI global, ??? (2010)

[45] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)

[46] Nguyen, Q.H., Ly, H.-B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I., Pham, B.T.: Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering **2021** (2021)

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work is supported by Istanbul Technical University (ITU) Vodafone Future Lab under Project Number 2018000463.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and materials

The dataset used in this study was requested from the original authors Hang Liu et al.[23]. It is not publicly available at the moment. However, we think the original authors will be willing to share the dataset upon request.

Author's Contribution

- Beltus Nkwawir Wiysobunri performed conceptualization, methodology, software, visualization, writing original draft.
- Hamza Salih Erden performed conceptualization, writing-review and editing, data acquisition supervision, and funding acquisition.
- Behcet Ugur Toreyin performed writing review and editing, validation, resources, and supervision.

Cover Letter

Dear Professor Antonio Di Nola;

We would like to submit the manuscript titled "CNN-based Server State Monitoring and Fault Diagnosis using Infrared Thermal Images" as an original research article in the Journal of Infrared Physics and Technology.

This article explores shallow and deep advanced convolutional neural network (CNN)-based architectures trained on server surface infrared thermal images for the automatic monitoring and diagnosis of five server operation conditions including faulty conditions.

The model architectures were evaluated using a stratified five-fold cross-validation resampling technique across five dataset split ratios. And experimental results indicated a significantly high predictive performance with accuracies greater than 98%.

Our experimental results outperformed those obtained using a traditional support vector machine algorithm by a maximum percentage point margin of 8.

This work is the first that has applied state-of-the-art deep pretrained CNN-based models and infrared thermography for the automatic diagnosis of the operation state of servers in data centers (DCs).

The application of this research study can pave the way towards the incorporation of non-contact infrared thermal imagers for the improvement DC thermal management, energy efficiency, and system reliability by DC operators. Therefore, it provides a benchmark for current advanced deep learning-based solutions for real-world problems.

This research study was inspired by the work of Liu et al. [23], who used signal processing and pattern recognition techniques together with traditional support vector machine algorithm to predict server status with an accuracy of 91.11 percent.

The proposed article also builds upon our previous preliminary work [29] with two shallow CNN models and one lightweight deep learning model to detect and classify server states using the thermal image dataset in the literature [23].

We declare this research article has not been published before nor submitted to any other journal for the consideration of publication. We kindly request your consideration, and we are looking forward to hearing from you.

Our primary objective was to submit this paper in response to the Soft Computing special issue call for papers in the Advances in Pattern Recognition and Computer Vision, Applications and Systems. However, this option was not listed under the Select article type section of the manuscript submission process. We hope that the required changes will be made on the website to rectify this problem.