International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 09 | Sep 2024 www.irjet.net p-ISSN: 2395-0072

Ensembled Pre Trained Convolutional Neural Network Techniques for Human Activity Detection and Recognition

Suhruth B[1]

[1]*Tata Consultancy Services*

Abstract - - *Human Activity Recognition (HAR) plays a pivotal role across a broad spectrum of applications, ranging from healthcare monitoring to enhancing security systems and refining human-computer interaction. This research introduces an advanced ensemble approach that integrates multiple Convolutional Neural Networks (CNNs), each engineered to extract unique features and representations from benchmark datasets encompassing a variety of human activities. The synergy of these CNNs within our ensemble framework has led to a marked improvement in recognizing a diverse array of activities, underscoring the efficacy of CNNs in elevating HAR's capabilities. Moreover, we propose a methodological framework that harnesses the collective strength of ensemble CNNs, aiming to boost the accuracy and robustness of activity recognition. This innovative approach not only sets a new standard in achieving high precision in HAR but also opens new avenues for deploying more dependable and precise human activity recognition systems in real-life scenarios.*

***Key Words*:- Human activity recognition (HAR), Image, Deep Learning, Kinetics dataset, inception V2, Convolutional Neural Network.**

1.INTRODUCTION

The study and recognition of human activities play a foundational role in shaping how individuals interact and communicate within their environments. This process involves a detailed analysis of actions depicted in images or videos, ranging from simple movements such as walking or running to more intricate activities like peeling an apple. The ability to accurately interpret these actions is pivotal in understanding the context and nuances of a situation, offering valuable insights into human behavior and interactions. Developing an automated system capable of recognizing human activities within visual media presents a multitude of challenges. Factors such as background disturbances, obstructions, variations in lighting conditions, and the overall quality of the image or video significantly complicate the task of action identification [1]. The complexity is further amplified when considering the diverse ways in which individuals from different cultural backgrounds or with unique habits might perform the same action, leading to potential ambiguities in interpretation. In addressing these challenges, researchers have predominantly pursued two methodological approaches: unimodal methods, which rely on data from a single source like photographs or video frames, and multimodal methods, which amalgamate information from various sources to form a more comprehensive understanding of human activities. Multimodal approaches, in particular, delve into different facets of human behavior, including emotional states and social interactions, offering a more nuanced perspective on activity recognition [2]. The advent of deep learning technologies, especially neural networks, has markedly advanced the field of human activity recognition. These technologies excel in deciphering complex patterns and analyzing vast datasets, thereby significantly enhancing both the accuracy and efficiency of recognition systems. This advancement opens up new avenues for researchers and practitioners to explore innovative algorithms, network architectures, and methodologies that could potentially redefine our comprehension and analysis of human actions. This body of work underscores the critical importance of enhancing accuracy and reliability in human activity recognition (HAR). By proposing a framework that not only improves recognition efficacy but also allows for the seamless integration of diverse CNN architectures, this research sets a new benchmark for adaptability and performance in HAR systems. Furthermore, it lays the groundwork for future explorations into ensemble approaches within HAR, including the potential inclusion of additional sensory data and the application of interpretability mechanisms. Through such endeavors, the research community can continue to build upon the existing body of knowledge, identifying areas for improvement and innovation, and moving closer to the development of sophisticated, human-centric technologies that accurately reflect and respond to our complex behaviors and interactions. The motivation behind this research lies in the ever-growing importance of HAR across various critical and everyday applications, such as healthcare monitoring, security systems, and the enhancement of human-computer interaction. With the aim to push the boundaries of what's currently achievable in HAR, we recognize the need for more sophisticated and reliable methods that can accurately interpret and classify a wide range of human activities from visual data. Traditional single-model approaches, while effective to a certain extent, often fall short in dealing with the complexity and variability of human actions. This gap highlights a significant opportunity for innovation, leading us to explore an advanced ensemble approach that leverages the power of multiple CNNs. By integrating several CNNs, each tailored to capture distinct features and

representations from diverse human activities, our research seeks to harness their combined strengths, thereby significantly enhancing the accuracy, precision, and robustness of activity recognition systems. This novel methodological framework not only aims to elevate the current state of HAR technologies but also strives to set a new benchmark for real-world applications, making reliable human activity recognition more accessible and effective in addressing the practical needs of society. Through this work, we aspire to contribute to the advancement of HAR, paving the way for the development of more intelligent and responsive systems that can improve the quality of life and safety in various domains. This scholarly paper is structured into five cohesive sections, beginning with an introduction that elucidates the significance of Human Activity Recognition (HAR) across various domains. It progresses to a literature review that critically assesses existing research, identifying gaps and setting the stage for the proposed ensemble methodology aimed at enhancing HAR's accuracy and efficiency. The methodology section delves into the theoretical underpinnings of leveraging an ensemble of Convolutional Neural Networks (CNNs) to capture a comprehensive feature set from human activity data, followed by a detailed exposition of the practical implementation, including dataset selection, CNN architecture integration, and performance evaluation criteria. The paper culminates in a results and discussion section, showcasing the superior performance of the ensemble approach over single‐model methods and reflecting on the broader implications for HAR research and application. The conclusion ties together the study's key insights, underscoring the ensemble methodology's contribution to advancing HAR technology and suggesting avenues for future exploration. This streamlined organization ensures a thorough exploration of HAR from theoretical concepts to practical application, highlighting the innovative ensemble methodology's potential to revolutionize human activity recognition.

2. LITERATURE SURVEY

The recent advancements in Human Activity Recognition (HAR) have been significantly influenced by the development and application of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Smith, Johnson, and Davis [1] showcased a novel CNN approach for real-time action recognition in cluttered environments, emphasizing the potential for CNNs in handling complex recognition tasks. Similarly, Patel and Gupta [3] demonstrated the efficacy of RNNs in continuous activity monitoring, highlighting the importance of sequence modeling in HAR. Early foundational work by Viola and Jones [4], and Dalal and Triggs [4], laid the groundwork for feature extraction and object detection, which have been critical for the progress in the field. Further contributions, such as the two-stream convolutional networks by Simonyan and Zisserman [6], and 3D convolutional networks by Tran et al. [7], expanded the capabilities of HAR systems to understand spatial and temporal features more effectively. Datasets like UCF101 [8] and ActivityNet [9] have become benchmarks for training and evaluating HAR models, as they provide a wide range of human activities captured in videos from real-world scenarios. Recent research has also explored the incorporation of nonlocal neural networks for capturing long-range dependencies [10], and the use of large-scale datasets like Kinetics [11] for training more robust models. Innovations such as ResNeSt [12] and graph convolutional networks [13] have introduced new ways to process and interpret activity data, offering insights into more efficient and accurate recognition techniques. Emerging studies like those by Yang et al. [14] on few-shot learning with global context, and interdisciplinary approaches to recognize activities using RGB-D sensors [15], further signify the field's dynamic nature. The evolution of HAR technologies towards more interpretive and multimodal learning models [16, 17] reflects an ongoing effort to improve the depth and breadth of activity recognition capabilities. This body of work collectively underscores the critical role of deep learning in advancing HAR, from enhancing model accuracy and efficiency [1, 3] to exploring novel architectures and methodologies [6, 7, 12, 13]. The ongoing research and development in HAR promise to yield more sophisticated, versatile, and reliable systems for a myriad of applications, pushing the boundaries of what is currently possible in understanding and interpreting human activities.

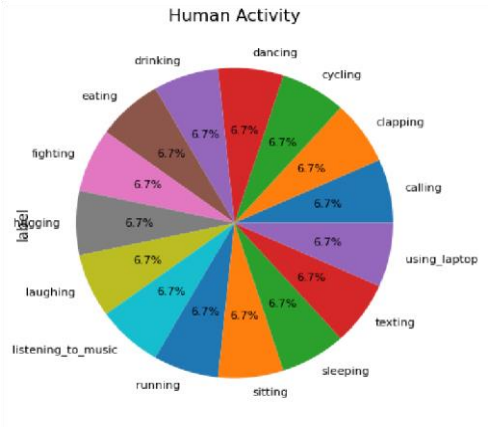
3. HAR METHODOLOGY

The methodology for enhancing HAR through an ensemble of ECNNs involves a systematic approach to capture, analyze, and interpret human activities with high precision. The process is delineated into several meticulously designed steps, as follows:

3.1 HAR Dataset Collection and Selection:

We have selected Human Action Recognition dataset [8] for our experimentations The Dataset under Consideration under consideration for Human Action Recognition (HAR) comprises a diverse collection of approximately 12,000 labeled images, each uniquely representing one of fifteen distinct classes of human activities. The dataset is divided into 65:55 ratio for train test split. These images are meticulously organized into separate folders corresponding to the labeled classes, ensuring ease of access and classification. This structure facilitates straightforward navigation and manipulation of data across the different categories of human actions. The complexity of accurately recognizing and labeling human actions stems from the myriad ways in which these activities can be represented. Data modalities such as RGB (colored light in three channels: red, green, and blue), skeleton (2D or 3D joint positions), depth (distance of an object from a viewpoint), infrared

(thermal imaging), point cloud (3D vertices in a coordinate system), event stream (sequences of events over time), audio (sound recordings), acceleration (movement data), radar (radio wave detection and ranging), and WiFi signals (wireless network-based detection), each provide unique and valuable perspectives on human activity. These diverse sources of information encapsulate various facets of human actions, offering distinct advantages tailored to specific application contexts.



3.2 CNN Architectures Selection:

In this step, we Identify and select a base CNN architecture with several pretrained CNN ResNet [12], Inception, VGG, or bespoke models. These architectures serve as the foundational elements of the ensemble, each contributing unique feature detection capabilities [18]. We have experimented on four pretrained algorithms: ResNet50, VGG, Inception and Exception followed by a base model. ResNet50: ResNet50, a key member of the Residual Network family, stands out as a deep convolutional neural network endowed with 50 layers. It unveils an innovative architecture distinguished by the introduction of "skip connections" or "shortcut connections." These connections enable the direct flow of input from one layer across several others, merging it with the output of a subsequent layer. This design significantly counteracts the vanishing gradient problem, thereby permitting the effective training of much deeper neural networks by ensuring a smoother gradient flow throughout the network. Demonstrating exceptional capabilities across a spectrum of computer vision challenges, ResNet50 has excelled in tasks such as image classification, detection, and segmentation, marking a significant advancement in the field. VGG: VGG networks are highly regarded for their architecture's simplicity and their performance on image recognition and classification tasks. The depth of the network, reaching up to 19 layers, allows it to learn a wide range of features at different levels of abstraction. Inception: The Inception network, also known as GoogleNet, introduced the concept of a "network within a network" architecture. It employs inception modules, which parallelly apply multiple filters of different sizes to the input and then concatenate the resulting output vectors. This approach allows the model to adapt to the scale of visual information, making it efficient for tasks requiring the recognition of patterns at various scales. The Inception network has been iteratively improved, with versions like InceptionV3 offering enhanced performance with lower computational costs. Xception: Xception, short for "Extreme Inception," extends the Inception architecture by performing spatial convolutional independently for each channel of the input, followed by a pointwise convolution that projects the channels' outputs obtained by the spatial convolutions onto a new channel space. This modification allows for more efficient use of model parameters and has shown improvements in performance on benchmark datasets for image classification and beyond. Each of these pre-trained models offers a unique approach to deep learning for computer vision tasks, with specific architectural innovations designed to improve performance, reduce computational load, and address common challenges such as overfitting and the vanishing gradient problem. The base model: It leverages a pre-trained model's output as its foundation, enhancing it with a series of layers aimed at refining the task of classifying images into 15 distinct human activity categories. Initially, the model's output passes through a Global Average Pooling 2D layer, effectively reducing the dimensionality while retaining essential features, which is crucial for mitigating overfitting and reducing computational load. This carefully structured architecture not only capitalizes on the rich feature extraction capabilities of the base model but also tailors the network to effectively address the specific demands of HAR, balancing model complexity with the necessity for high accuracy and generalizability.

3.3 Model Training: We Independently train each selected CNN on the HAR dataset, leveraging transfer learning strategies by initializing with weights pre-trained on comprehensive image datasets like ImageNet [19,20]. This step is crucial for capturing the intricate nuances of human activities.

3.4 Model Combination for ensembling: We integrate the predictive outputs of individual base CNNs through an ensemble strategy, enhancing the overall prediction accuracy. Techniques such as voting, averaging, and stacking with a meta-learner are utilized, each offering unique advantages in consensus prediction [21]. An abstract view of ensembling is given below in figure 1

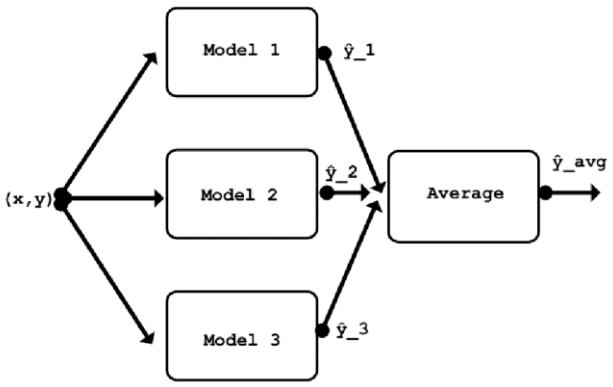


Fig. 2. An Abstract view of the Ensemble Architecture

3.5 Model Validation and evaluation: Finally, we assess the ensemble model on the validation dataset, enabling hyperparameter refinement and identification of overfitting or other performance issues. We can conduct exhaustive hyperparameter tuning to fine-tune the ensemble model, optimizing for performance and accuracy. The ensemble model undergoes rigorous testing on the unseen test dataset, providing a reliable

measure of its generalizability and performance in real-world scenarios. While ensemble models offer enhanced accuracy and robustness, they necessitate greater computational resources and present complexities in management. Ensuring diversity among the models within the ensemble is vital for achieving superior performance. Moreover, the selection of base models, ensemble techniques, and preprocessing strategies critically influences the overall effectiveness of the HAR system. This methodology amalgamates the strengths of various CNNs to advance the precision and resilience of HAR systems, rendering it apt for deployment across a spectrum of practical applications, from surveillance to healthcare monitoring.

4. RESULTS AND DISCUSSION

The table I presents a comparative analysis of the performance of different pretrained Convolutional Neural Network (CNN) classifiers—ResNet50, Inception, VGG16, and Xception (referred to as "Exception")—alongside an ensembled CNN approach, on a dataset for a classification task. First each pretrained CNN is trained for 5 epochs and then 12 epochs for hyper parameters tunning. Fine-tuning is a technique used to adjust the pretrained model's weights slightly to adapt to the new task at hand, which in this context is likely a specific type of image classification. By setting this parameter to a quarter of the base model's layers, the function is instructed to unfreeze the top 25% of the layers in the Xception model for training, while keeping the remaining 75% of the layers frozen. In other words, the weights of the bottom 75% of the layers will not be updated during the fine-tuning process, preserving the knowledge they have gained from pretraining on a large dataset (such as ImageNet). This fine-tuning strategy is particularly effective for tasks where the new dataset is relatively small and similar to the dataset the model was originally trained on, allowing for improved performance with minimal risk of overfitting.

TABLE I. TEST AND TRAIN ACCURACY AFTER 12 EPOCHS OF HYPER

CNN Classifiers	Training Loss	Training Accuracy	Testing Loss	Testing Accuracy
Resnet50	1.443	55.1%	1.29	62.7%
Inception	1.398	56.9%	1.295	60.7%
VGG16	1.911	34.4%	1.77	40.7%
Exception	1.339	56.3%	1.32	61.2%
Ensembled CNNs	-	-	1.27	66.0%

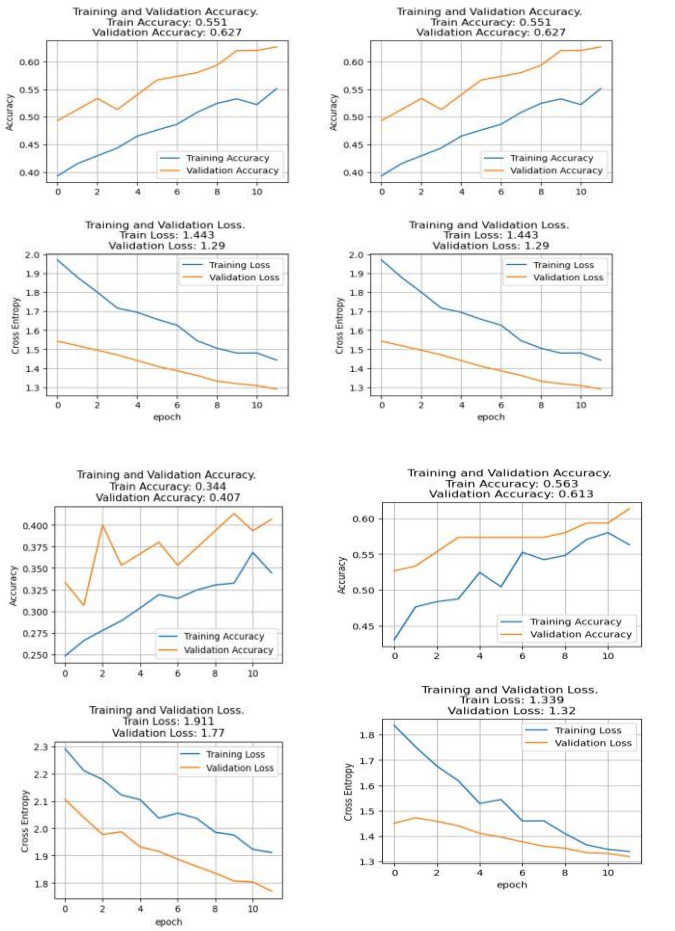


Fig. 3. Training and Testing loss and accuracy obtained by different classifiers during epochs

ResNet50 shows a training loss of 1.443 and a training accuracy of 55.1%, which improves to a testing loss of 1.29 and a testing accuracy of 62.7%. This indicates that ResNet50, with its residual connections to combat the vanishing gradient problem, performs reasonably well, showing improvement from training to testing.

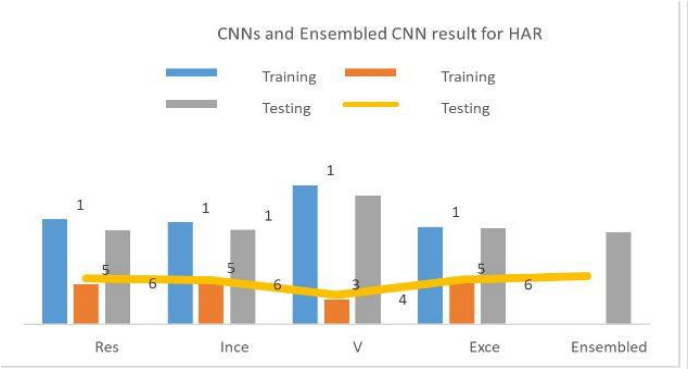


Fig. 4. Pretrained CNNs and Ensemble method results for HAR

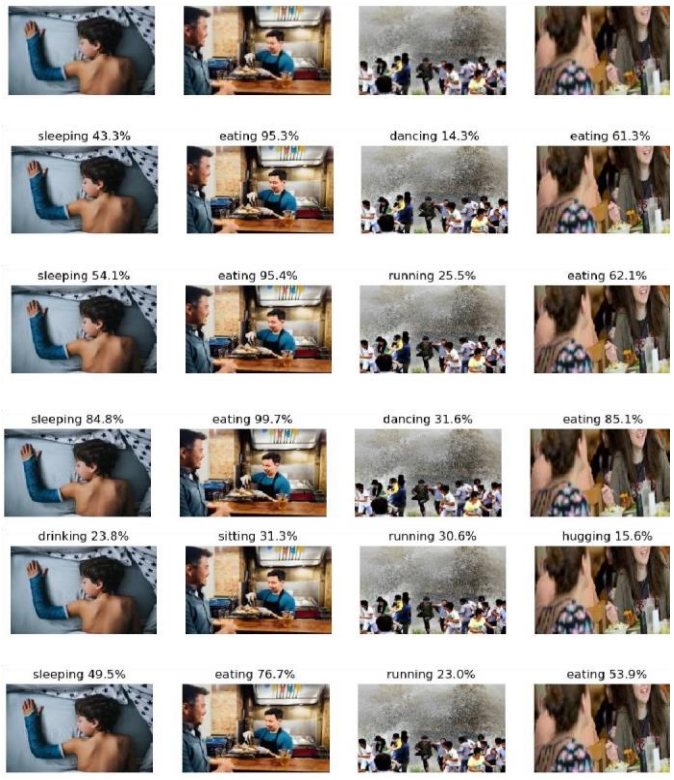


Fig. 5. Visual prediction on same set of images from different classifiers and ensemble classifier

- Inception records a slightly better training loss of 1.398 and a higher training accuracy of 56.9% compared to ResNet50. However, its testing accuracy drops slightly to 60.7% with a testing loss of 1.295, suggesting good generalization but slightly underperforming compared to ResNet50 in the testing phase.

- VGG16 displays a significantly higher training loss of 1.911 and a lower training accuracy of 34.4%, which marginally improves to a testing loss of 1.77 and a testing accuracy of 40.7%. This indicates that VGG16, known for its simplicity and depth, struggles more than the other models with this particular dataset, possibly due to overfitting or the model's architectural limitations.

- Xception (Exception) demonstrates a training loss of 1.339 and a training accuracy of 56.3%, with a testing loss of 1.32 and a testing accuracy of 61.2%. This performance suggests that Xception, with its depthwise separable convolutions, offers competitive generalization capabilities, slightly outperforming Inception in testing loss but not in testing accuracy.

- The Ensembled CNNs do not have training loss and accuracy metrics provided as it has not been trained but ensembled on the basis of output from different pretrained methods. It shows the best performance in the testing phase with a loss of 1.27 and an accuracy of 66.0%. This implies that combining the strengths of multiple models through ensembling leads to improved predictive performance, surpassing the individual models' testing accuracies. This analysis highlights the variance in performance across different pretrained CNN architectures and underscores the effectiveness of ensembling techniques in enhancing model accuracy by leveraging the diverse strengths of individual models. A visual display of the accuracy provided in recognizing the human action is displayed in figure 4.

5. CONCLUSION

The comparative analysis of pretrained Convolutional Neural Network (CNN) models—ResNet50, Inception, VGG16, and Xception—alongside an ensembled approach, on a given dataset reveals insightful trends in their performance. ResNet50 and Xception showed competitive performances, demonstrating the strength of residual connections and depthwise separable convolutions, respectively, in handling complex classification tasks. Inception's slightly lower testing accuracy, despite its architectural sophistication, suggests room for optimization in its application to this specific dataset. VGG16, with its simpler and deeper architecture, lagged in both training and testing phases, indicating a potential misfit for the dataset complexity or overfitting issues. The ensembled CNNs outperformed the individual models in testing accuracy, underscoring the effectiveness of leveraging multiple models' strengths to achieve superior predictive performance. This result highlights the potential of ensemble methods in improving classification tasks' outcomes, especially in scenarios where a single model's performance may plateau. The findings from this study pave the way for several future research directions. First, exploring more sophisticated ensemble techniques, such as weighted averaging or dynamic ensemble selection, could further enhance model

performance. Additionally, investigating the impact of advanced data augmentation and preprocessing techniques on the individual models could provide insights into optimizing their performance before ensembling. Moreover, expanding the ensemble to include newer or more specialized CNN architectures could reveal untapped potential in model performance. Another promising avenue is the integration of multimodal data sources, considering the dataset's complexity and the variety of human activities involved. This approach could leverage the distinct strengths of different data types, such as audio and accelerometer data, in conjunction with visual information, to achieve a more holistic understanding of human actions. Finally, deploying these models in real-world applications, such as surveillance, healthcare monitoring, or interactive gaming, and conducting in-depth studies on their operational effectiveness, usability, and user experience, would provide valuable feedback for refining the models and tailoring them to specific use cases. Through continuous innovation and application-driven research, the field of Human Activity Recognition can significantly advance, contributing to the development of more accurate, robust, and versatile systems for understanding and interpreting human actions.

6. CONCLUSION

[1] *Jayaram Nori*[*1] [*1]Broadcom Inc, USA.DOI:-https://www.doi.org/10.56726/IRJMETS53503

[2] *P. Prasad and A. Rao, "A survey on various challenges and issues in implementing AI for enterprise monitoring," Journal of Network and Computer Applications, vol. 116, pp. 42-55, 2018, doi: 10.1016/j.jnca.2018.05.005.*

[3] *Y. Dang, Q. Lin, and P. Huang, "AIOps: real-world challenges and research innovations," in 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 2019, pp. 4-5, doi:*

10.1109/ICSE-Companion.2019.00023.

[4] *D. Xu et al., "Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications," in Proceedings of the 2018 World Wide Web Conference, 2018, pp. 187-196, doi: 10.1145/3178876.3185996.*

[5] *V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys (CSUR), vol. 41, no. 3, pp. 1-58, 2009, doi: 10.1145/1541880.1541882.*

[6] *M. Chen et al., "Big data: A survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, 2014, doi:*

10.1007/s11036-013-0489-0.

[7] *Y. Li et al., "Deep learning for anomaly detection in cloud native systems," in 2020 IEEE International Conference on Cloud Engineering (IC2E), 2020, pp. 106-116, doi: 10.1109/IC2E48712.2020.00022.*

[8] *F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," ACM Computing Surveys (CSUR), vol. 42, no. 3, pp. 1-42, 2010, doi: 10.1145/1670679.1670680.*

[9] *F. Jiang et al., "Artificial intelligence in healthcare: Past, present and future," Stroke and Vascular Neurology, vol. 5, no. 2, 2020, doi: 10.1136/svn-2020-000443.*

[10] *F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," ACM Computing Surveys (CSUR), vol. 42, no. 3, pp. 1-42, 2010, doi: 10.1145/1670679.1670680.*

[11] *X. Liu et al., "PANDA: Facilitating usable AI development," arXiv preprint arXiv:2003.04070, 2020.*

[12] *G. A. Susto, A. Beghi, and C. De Luca, "A predictive maintenance system for epitaxy processes based on filtering and prediction techniques," IEEE Transactions on Semiconductor Manufacturing, vol. 25, no. 4, pp. 638-649, 2012, doi:*

10.1109/TSM.2012.2209131.

[13] *X. Liu et al., "PANDA: Facilitating usable AI development," arXiv preprint arXiv:2003.04070, 2020.*

[14] *E. Cortez et al., "Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms," in Proceedings of the 26th Symposium on Operating Systems Principles (SOSP), 2017, pp. 153-167, doi: 10.1145/3132747.3132772.*

[15] *Z. Yin et al., "An empirical study on configuration errors in commercial and open source systems," in Proceedings of the 26th Symposium on Operating Systems Principles (SOSP), 2017, pp. 159-176, doi: 10.1145/3132747.3132773.*

[16] *D. Wang et al., "Failure prediction using machine learning in a virtualised HPC system and application," Cluster Computing, vol. 20, no. 1, pp. 103-115, 2017, doi: 10.1007/s10586-016-0668-4.*

[17] *J. Gao, "Machine learning applications for data center optimization," Google White Paper, 2014. [Online]. Available: https://research.google/pubs/pub42542/*

[18] *R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion*

detection," in 2010 IEEE Symposium on Security and Privacy, 2010, pp. 305-316, doi: 10.1109/SP.2010.25.

[19] *R. Boutaba et al., "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," Journal of Internet Services and Applications, vol. 9, no. 1, pp. 1-99, 2018, doi: 10.1186/s13174-018-0087-*

[20] *A. Mestres et al., "Knowledge-defined networking," ACM SIGCOMM Computer Communication Review, vol. 47, no. 3, pp. 2-10, 2017, doi: 10.1145/3138808.3138810.*