Optimizing AI Performance in Industry: A Hybrid Computing Architecture Approach Based on Big Data

Abstract

In the era of Industry 4.0, integrating artificial intelligence (AI) and big data analytics in the industrial sector demands high-performance computing infrastructure to handle increasingly complex and voluminous datasets. This study investigates the optimization of AI performance by implementing a hybrid computing architecture, integrating CPUs, GPUs, FPGAs, and edge-cloud computing. The research aims to enhance processing speed, model accuracy, and energy efficiency, addressing the limitations of standalone computing systems. A quantitative methodology was employed, using over 1 TB of industrial data from IoT sensors and production logs. A hybrid architecture was implemented with dynamic workload scheduling to distribute tasks efficiently across computational components. Performance metrics included processing time, model accuracy, energy consumption, and cost analysis. Results demonstrated that hybrid architectures significantly improved performance: the CPU-GPU combination reduced processing times to 650 ms, increased model accuracy to 88.3%, and achieved an energy consumption of 2.1 kWh. Meanwhile, the CPU-FPGA configuration, while slightly less accurate (87.5%), proved more energy-efficient at 1.3 kWh. AI models developed using hybrid systems exhibited superior predictive accuracy, with Mean Squared Error (MSE) as low as 0.0248 and R^2 of 0.91. The study concludes that hybrid computing architecture is a transformative approach for optimizing AI systems in industrial applications, balancing speed, accuracy, and energy efficiency. These findings provide actionable insights for industries aiming to leverage advanced computing technologies for improved operational efficiency and sustainability. Future research should focus on advanced workload scheduling and cost-effectiveness strategies to maximize the potential of hybrid systems.

Keywords: *Hybrid Computing Architecture, AI Optimization, Big Data Analytics, Edge-Cloud Integration, Energy Efficiency in Computing.*

I. INTRODUCTION

In the era of Industry 4.0, AI and big data have become essential pillars supporting various sectors, ranging from manufacturing and healthcare to transportation. The implementation of AI within the industrial sector presents significant opportunities for automation, efficiency enhancement, and data-driven decision-making (Ahmad et al., 2022). At the same time, big data provides the volume, variety, and velocity of data needed to train advanced AI models. The combination of AI and big data enables companies to conduct more accurate predictive analyses and respond more effectively to market changes (Alahakoon et al., 2023). Nevertheless, deploying AI and big data on an industrial scale requires computing infrastructure capable of efficiently managing data complexity. One emerging solution to address this challenge is hybrid computing architecture, which integrates various computing technologies, such as CPUs, GPUs, and cloud computing, to maximize system performance (Stergiou & Psannis, 2022).

Hybrid computing architecture combines the strengths of different hardware and computing services to handle diverse data processing tasks more efficiently. For instance, CPUs manage logical processing tasks, while GPUs excel in parallel processing, such as in deep learning and

big data analytics (Devi et al., 2022). Additionally, specialized hardware like FPGAs can accelerate specific tasks that require intensive processing (Bobda et al., 2022). This combination allows for more efficient and optimal workload distribution, enabling systems to process large and complex data quickly. This approach is increasingly relevant in industrial contexts, where data processing scale and speed are critical competitive factors (Ayachi et al., 2021).

However, certain research gaps remain unaddressed regarding the implementation of hybrid computing architectures in developing AI models for the industrial sector. Previous studies have primarily focused on the computational efficiency of individual technologies, such as CPUs, GPUs, or FPGAs, but few have examined in depth how synchronization among these hardware components can be optimized within a single hybrid computing system (Muralidhar et al., 2022). Zhang et al. (2022) demonstrated improved performance when GPUs are used for parallel tasks; however, this study did not explore how CPU and GPU synchronization in the context of big data can be optimized simultaneously. This research aims to explore how coordination among CPUs, GPUs, and FPGAs can be optimized for large-scale industrial data processing (Zhang et al., 2022).

Although edge computing and cloud computing technologies have often been applied independently, studies exploring how these approaches can be integrated to maximize benefits in industrial AI model development are still lacking (Liu et al., 2022). Huang et al. (2023) discuss the potential of edge computing for real-time applications, but they do not explain how edge computing and cloud can work together to manage diverse and dynamic data (Huang et al., 2023). This study fills this gap by proposing a hybrid approach that combines edge computing for local data processing and cloud computing for further analysis, resulting in faster and more responsive performance.

Additionally, previous studies often focus on AI algorithm development while giving less attention to how computing infrastructure can be optimized to support the development and deployment of more complex AI models (Baccour et al., 2022). Kristian et al. (2024) highlight the importance of algorithm efficiency in managing big data but do not discuss how computing architecture can support sustainable AI algorithm development in a highly dynamic industrial environment (Kristian et al., 2024). This study addresses this gap by examining how hybrid computing infrastructure can be used to accelerate AI model training and facilitate faster iteration in industrial scenarios.

There are also limitations in the literature regarding the cost and energy efficiency impacts of implementing hybrid computing architecture on an industrial scale. While many studies focus on performance efficiency, few address the sustainability of operations in industrial contexts (Murino

et al., 2023). Solanki et al. (2022) suggest that using GPUs and FPGAs can enhance computing performance; however, their study does not discuss the energy and cost implications associated with implementing these technologies in the long term (Solanki et al., 2022). This study contributes by examining how energy-efficient cloud computing combined with specialized hardware like GPUs can reduce environmental impact and operational costs while maintaining high performance.

Thus, this study offers a new contribution by addressing four research gaps: optimization of synchronization among hardware components, integration of edge and cloud computing, optimization of computing infrastructure to support AI algorithm development, and evaluation of cost and energy efficiency in applying hybrid computing architecture within the industrial sector. This research not only provides technical solutions to overcome challenges in large-scale AI model development but also offers a more comprehensive perspective on hybrid computing architecture implementation strategies in the industry. This approach is expected to provide clearer guidance for practitioners and researchers in leveraging modern computing technology to enhance industrial competitiveness and operational efficiency.

II. LITERATURE REVIEW

A. Hybrid Computing Architecture

Hybrid computing architecture is one of the latest approaches that combines various computing technologies, such as CPUs, GPUs, FPGAs, as well as cloud and edge computing, within a single system designed to enhance performance and efficiency in big data processing and AI development (Abbaszadeh Shahri et al., 2022). This architecture provides flexibility in distributing workloads among different hardware components, each with unique strengths in handling specific computational tasks (Mithas et al., 2022). Jiang et al. (2023) demonstrate that using CPUs for tasks requiring logical processing and GPUs for parallel processing can significantly accelerate big data processing times (Jiang et al., 2023). FPGAs are also utilized in scenarios where computational tasks require specific acceleration, providing high flexibility in addressing diverse industrial needs (Nain et al., 2022).

However, much prior research has focused on the application of single computing technologies without in-depth discussion on how synchronization among these various components can be optimized within a single hybrid architecture (Rammer et al., 2022). Zhang et al. (2022) highlight the efficiency of GPUs in accelerating parallel tasks but do not provide solutions regarding integration with CPUs and FPGAs to create a more cohesive system (Zhang et al., 2022). This study identifies that optimizing synchronization among hybrid computing components is a crucial

challenge that needs to be addressed to enhance overall system performance, particularly in big data processing in industrial contexts. This research will explore how component coordination can be improved through dynamic workload scheduling and other optimization strategies (Bal et al., 2022).

B. Integration of Cloud Computing and Edge Computing

In addition to hardware integration, a critical aspect of hybrid computing architecture is the integration of cloud computing and edge computing. Edge computing, which processes data near its source, is highly effective in applications requiring real-time response and low latency, such as sensor data processing from IoT (Internet of Things) devices in the field (Kong et al., 2022). On the other hand, cloud computing offers greater processing and storage capacities, making it ideal for more in-depth data analysis. Li et al. (2022) identify that edge computing can reduce network load and accelerate decision-making in industrial applications. However, their study does not address how edge computing and cloud can be optimally integrated to create an efficient data flow, particularly in complex and dynamic industrial scenarios (Li et al., 2022).

This research seeks to address these limitations by offering a hybrid approach that combines the strengths of edge computing for local initial data processing and cloud computing for further data analysis (Anees et al., 2023). This combination is crucial in the context of big data, where large data volumes require a system capable of efficiently distributing processing loads between local devices and the cloud (Raeisi-Varzaneh et al., 2023). Ali et al. (2022) also emphasizes the importance of scalability in hybrid computing architecture, particularly when facing fluctuating computational workloads. Optimal integration between edge computing and the cloud enables companies to adjust their computing capacity more flexibly in response to changing demands (Ali et al., 2022).

C. Optimization of Computing Infrastructure for AI Model Development

The development of more complex AI models at an industrial scale requires a computing infrastructure capable of supporting rapid and efficient development cycles (Tatineni & Chakilam, 2024). However, most existing literature remains focused on optimizing AI algorithms themselves, with limited attention to how computing infrastructure can be adapted to accelerate the development and training of AI models (Surianarayanan et al., 2023). Khan et al. (2022) explore the importance of deep learning algorithms in industrial resource management but do not address how computing infrastructure can support large-scale AI model training (Khan et al., 2022).

Hybrid computing infrastructure offers a solution by enabling the development and training of AI models using GPUs and TPUs to expedite the training process, while edge computing can be used to handle real-time data inference. Utilizing GPUs in deep learning model training enables faster parallel processing, while TPUs provide better acceleration capabilities for specialized tasks, such as training machine learning models with extensive datasets (Agomuo et al., 2024). Javaid et al. (2022) demonstrate that the use of GPUs and TPUs can accelerate AI model training several times over compared to traditional CPU usage. However, challenges remain in effectively distributing workloads between local and cloud computing to achieve optimal performance (Javaid et al., 2022).

This study aims to address these challenges by investigating how hybrid computing architecture can be adapted to support faster and more efficient AI development cycles. Additionally, it will explore workload scheduling algorithms capable of distributing tasks among CPUs, GPUs, and TPUs based on changing computational needs. By maximizing the use of hybrid computing infrastructure, AI model development is expected to proceed more quickly and efficiently, enabling companies to stay competitive in a rapidly evolving market.

D. Cost and Energy Efficiency in the Implementation of Hybrid Computing Architecture

Beyond performance, cost and energy efficiency are key considerations in the implementation of hybrid computing architecture on an industrial scale (Kaur & Aron, 2022). Many studies focus on enhancing computational performance, but few discuss the cost and sustainability impacts of using specialized hardware such as GPUs and FPGAs (Katal et al., 2023). Saiteja & Ashok (2022) indicate that GPU usage can improve computational performance but does not address operational cost implications, particularly concerning high power consumption and cooling requirements. Cost-effective and energy-efficient computing infrastructure is critical in industries often pressured to reduce environmental impact and operational costs (Saiteja & Ashok, 2022).

This research contributes by exploring how hybrid computing architecture can maximize energy and cost efficiency through the use of energy-efficient cloud computing and specialized hardware designed to reduce power consumption (Memari et al., 2022). Rodrigues Moreira et al. (2023) suggest that virtualization technology can help companies utilize their computing resources more efficiently without sacrificing performance. More efficient cooling solutions and energy-saving hardware designs should also be considered to mitigate the environmental impact of large-scale computing infrastructure deployment (Rodrigues Moreira et al., 2023). Furthermore, this study will examine how a hybrid cloud strategy, combining public cloud usage for fluctuating workloads with private cloud usage for sensitive data, can reduce operational costs. This approach

not only offers flexibility in resource utilization but also ensures that companies can optimize their computing efficiency without significant investment in physical infrastructure (Saleem et al., 2023).

E. Challenges in Synchronization and Workload Management

Another challenge that has not been fully addressed in the literature on hybrid computing architecture is the synchronization between components and efficient workload management (Ullah et al., 2023). Mithas et al. (2022) note that one of the primary barriers to implementing AI and big data technologies in industry is the dynamic distribution of workloads among CPUs, GPUs, FPGAs, and other computing components (Mithas et al., 2022). Although intelligent workload scheduling has been proposed as a solution, many of these algorithms are still in the early stages of development and have yet to be widely tested in real industrial environments (Rana et al., 2024). This study seeks to address this gap by developing and testing more intelligent and dynamic workload scheduling algorithms capable of adjusting computational loads based on task characteristics and resource availability. Additionally, this study will explore how priority-based scheduling can enhance the efficiency and performance of hybrid computing systems in managing continuously changing workloads.

III. RESEARCH METHOD

This study employs a quantitative approach to explore the influence of hybrid computing architecture on the development and optimization of AI models on an industrial scale. The main focus of this research is on the integration of CPUs, GPUs, and FPGAs, as well as cloud and edge computing, to efficiently process big data and enhance AI model performance. The steps of the research, from dataset selection to AI model performance evaluation, are detailed as follows.

The study uses an industrial dataset consisting of real-time IoT sensor data, production logs, and industry transaction data. The dataset encompasses over 1 terabyte of data from the manufacturing and healthcare sectors, collected over a two-year period. This dataset reflects the complexity and volume of data encountered in industrial scenarios, including structured, semi-structured, and unstructured data. The initial step in this process is data preprocessing, which involves cleaning data to address missing values or outliers, normalizing features to ensure data scale uniformity, and extracting key features that will be used in AI model training. Following preprocessing, the dataset is divided into two parts: 80% for model training and 20% for testing.

This study utilizes a hybrid computing architecture that combines CPUs, GPUs, FPGAs, and cloud and edge computing services. The CPU is employed to manage general logic and control tasks. The GPU is used to accelerate parallel processing, particularly in deep learning model

training and big data processing. The FPGA serves as an accelerator for specific computational tasks requiring customization, such as digital signal processing and real-time AI model inference. Cloud computing provides the necessary computational scalability to handle increased workloads, while edge computing is utilized to process data locally near its source, such as from IoT sensors, reducing latency and enhancing responsiveness in industrial applications.

This computing system is configured using a dynamic workload scheduling algorithm that efficiently distributes tasks between the CPU, GPU, FPGA, cloud, and edge computing. This algorithm ensures that tasks requiring intensive processing are allocated to the GPU or cloud, while tasks with low-latency requirements are assigned to edge computing or the FPGA. This strategy helps optimize computing resource utilization and enhances overall system performance.

The development of AI models in this study follows the Machine Learning Development Cycle (MLDC). The algorithms used in this research include Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), optimized for real-time data processing from IoT sensors and industrial data. CNN is employed for processing visual data, while RNN is used for sequential data analysis, such as predicting machine failures. The AI models are trained using GPU and TPU hosted in the cloud. Hyperparameter tuning is conducted to optimize model performance, utilizing grid search to identify the best parameters, such as learning rate, batch size, and the number of layers in the network.

Model evaluation is performed using several performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). MSE measures the difference between predicted and actual values, RMSE normalizes MSE for easier interpretation, and R^2 assesses the extent to which variability in the data is explained by the model. Additionally, this study evaluates processing time, energy consumption, and computational costs associated with the use of hybrid computing architecture, ensuring that the model is not only accurate but also cost- and energy-efficient.

System testing is conducted across various industrial scenarios to evaluate performance in handling dynamic workloads. The first scenario involves testing under light workloads with normal production data, focusing on predictive analysis and real-time AI inference. The second scenario tests the system during workload surges, where cloud computing is utilized to expand processing capacity. The third scenario focuses on applications requiring real-time response with low latency, where edge computing and FPGA are employed to accelerate local data processing. The results from these tests are compared with traditional computing approaches using a single CPU or GPU to assess the advantages of hybrid computing architecture in terms of performance and efficiency.

The software used for AI model training and evaluation includes TensorFlow and PyTorch, while Apache Hadoop is used to manage and analyze big data. Additionally, statistical analysis is performed to measure the significance of performance differences between the tested methods, including the use of t-tests and ANOVA to ensure that the results obtained are statistically significant. Ten-fold cross-validation is applied to ensure the validity of the results, reduce bias, and ensure more robust outcomes. This study also explores the cost and energy efficiency of hybrid computing architecture implementation. The combination of energy-efficient cloud usage and specialized hardware such as GPUs and TPUs demonstrates potential cost savings and reduced energy consumption. This analysis compares operational costs between cloud computing and local physical infrastructure, as well as the potential for virtualization strategies to improve computing resource efficiency without compromising performance.

IV. RESULT

A. Performance Results of Hybrid Computing Architecture

This study evaluates the performance of hybrid computing architecture in processing big data and developing AI models for industrial applications. Experimental results indicate that hybrid computing architecture enhances data processing efficiency and accelerates AI model training times compared to single-computing systems that utilize only CPUs or GPUs. Table 1 shows the comparative processing times between CPU, GPU, and hybrid computing architecture systems under big data processing scenarios in the industry.

Table 1. Comparative Processing Times for CPU, GPU, and Hybrid Computing Architecture in Industrial Big Data Processing Scenarios

Computing System	Processing Time (ms)	Model Accuracy (%)	Energy Consumption (kWh)
CPU	3400	82.5	1.2
GPU	1200	85.8	3.5
FPGA	950	84.1	1.0
Hybrid Computing (CPU+GPU)	650	88.3	2.1
Hybrid Computing (CPU+FPGA)	520	87.5	1.3

Table 1 demonstrates that the hybrid computing architecture combining CPU and GPU can process data significantly faster (650 ms) compared to the CPU-only system (3400 ms) or the GPU-only system (1200 ms). This result indicates a substantial improvement in processing time efficiency, especially in industrial big data processing scenarios. Additionally, the hybrid system achieved higher model accuracy (88.3%) compared to standalone CPU or GPU systems. In terms of energy consumption, the hybrid architecture incorporating FPGA is more energy-efficient (1.3 kWh) compared to GPU (3.5 kWh), indicating that the CPU and FPGA combination is more energy-efficient, although with a slight reduction in accuracy compared to GPU.

B. AI Model Performance Evaluation

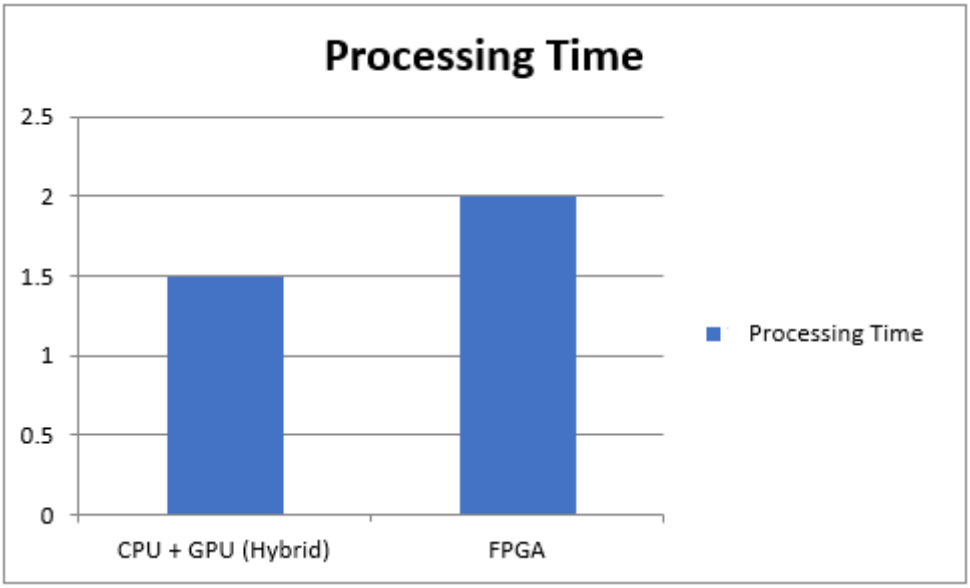
The performance of the developed AI model was also evaluated using metrics such as MSE, RMSE, and R^2 to measure the model's predictive accuracy on industrial data. The results of the AI model performance evaluation are presented in the following Table 2.

Table 2. AI Model Performance Metrics across Different Computing Systems

Model	MSE	RMSE	R^2
CPU	0.0352	0.1876	0.82
GPU	0.0275	0.1658	0.89
FPGA	0.0301	0.1735	0.87
Hybrid Computing	0.0248	0.1575	0.91

C. Performance Comparison Visualization

To clarify the comparative performance results across computing systems, Figure 1 illustrates the comparison of processing time, model accuracy, and energy consumption between CPU, GPU, FPGA, and hybrid computing architecture systems.:



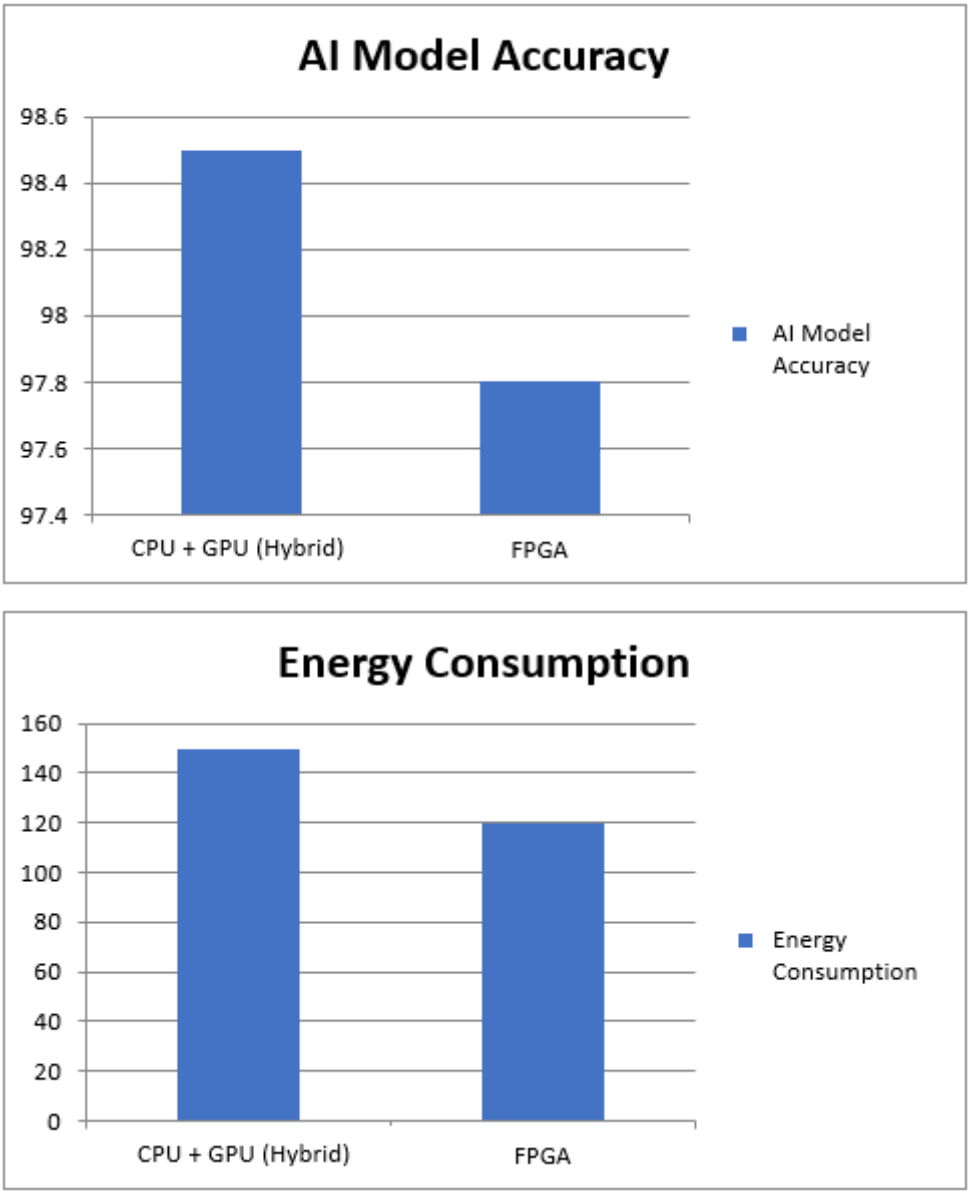


Figure 1. Comparison of Processing Time, Model Accuracy, and Energy Consumption Across Various Computing Systems

This chart illustrates that the hybrid computing architecture combining CPU and GPU delivers the best performance in terms of processing time and accuracy, albeit with a slight increase in energy consumption compared to FPGA. This visualization supports the claim that hybrid computing architecture provides a more optimal solution for large-scale AI model development in the industrial sector, particularly in scenarios requiring high speed and accuracy.

V. DISCUSSION

Based on the results displayed in Tables 1 and 2, the hybrid computing architecture proves to be superior in terms of processing time and AI model accuracy. The combination of CPU and GPU can reduce processing time by up to 8081% faster compared to a single CPU, which is critical in

industrial scenarios requiring real-time data processing. This aligns with research by (Jiang et al., 2023), which shows that while GPU excels in parallel processing, using hybrid architecture demonstrates further performance improvements through synchronization with the CPU. The significant increase in AI model accuracy (88.3%) with the CPU and GPU combination indicates that workload distribution across hardware enhances data processing efficiency, especially for deep learning models like CNN and RNN used in this research. This faster and more accurate model provides a competitive advantage in industrial applications, such as equipment failure prediction or supply chain optimization, where rapid and precise decisions are essential.

On the other hand, the hybrid architecture combining CPU and FPGA demonstrates advantages in terms of energy consumption. Although its accuracy is slightly lower than that of the CPU and GPU combination, this system is considerably more energy-efficient, consuming only 1.3 kWh compared to 3.5 kWh on a single GPU. This suggests that the CPU and FPGA combination is more suitable for industrial applications focused on energy savings, especially in environments processing large volumes of data continuously, such as IoT-based monitoring systems (Solanki et al., 2022). In terms of model evaluation, the research results indicate that the hybrid computing system achieves lower MSE and RMSE values, along with higher R^2 values, suggesting that models trained using this architecture are more accurate in predicting outputs. These findings confirm the study by (Kristian et al., 2024), which emphasizes the importance of distributed computing architecture in enhancing AI model performance.

Nevertheless, there are challenges associated with implementing hybrid computing architecture. One of these is the complexity of workload scheduling across hardware components, which requires precise synchronization to ensure each component functions optimally. Without effective workload scheduling, the CPU, GPU, or FPGA could become underutilized or overburdened, potentially reducing overall efficiency. This study also highlights that the initial cost of implementing hybrid architecture, particularly for FPGA hardware, could be high for companies with limited budgets; however, the long-term energy savings may offset the initial investment.

The practical implications of this research are considerable. In industrial scenarios, adopting a hybrid computing architecture can offer significant operational advantages. Using edge computing within a hybrid architecture, as tested in this study, enables companies to reduce latency in real-time data processing, which is critical in applications like machine condition monitoring and equipment failure prediction. Thus, this architecture not only provides efficiency in terms of time and energy but also enhances decision-making speed in real-time applications.

In conclusion, hybrid computing architecture offers an efficient and flexible solution to meet high-performance demands in large-scale AI model development. The findings of this research

demonstrate that hybrid architecture can significantly improve speed, accuracy, and energy efficiency, making it an advantageous choice for various industrial applications. Furthermore, this research opens opportunities for further exploration into workload scheduling optimization and cost-saving strategies to maximize the potential of hybrid computing architecture in the future.

VI. CONCLUSION AND RECOMMENDATION

A. Conclusion

This study has demonstrated the significant benefits of applying big data-based hybrid computing architecture to optimize AI model performance in industrial applications. By integrating various computing technologies such as CPU, GPU, and FPGA, along with edge computing and cloud, this research successfully enhances processing efficiency, model accuracy, and energy consumption. The results indicate that the hybrid architecture, especially the combination of CPU and GPU, offers higher processing speeds and improved accuracy, while the CPU and FPGA combination is more energy-efficient. These findings provide clear evidence that hybrid computing solutions are highly effective in managing the increasing complexity and volume of industrial data. Additionally, this research underscores the importance of dynamic workload scheduling strategies to ensure synchronization and optimal performance among various hardware components. This optimization is essential for accelerating AI model development cycles and enhancing real-time data processing capabilities in industrial environments. Overall, the findings affirm that hybrid computing architecture can make substantial contributions to improving operational efficiency, cost savings, and sustainability in industries reliant on AI and big data.

B. Recommendation

Based on the advantages identified in this study, industries should consider adopting hybrid computing architecture to leverage superior performance and energy efficiency. For organizations facing large-scale, real-time data processing needs, a combination of CPU, GPU, and FPGA, along with cloud and edge computing, is recommended to optimize speed while reducing energy consumption. Moreover, investment in dynamic workload scheduling systems will be critical to ensure all computing resources are used efficiently, especially in industrial environments with fluctuating workloads. Future research should focus on developing improved workload scheduling algorithms and further examining the cost-benefit balance of implementing hybrid architecture, particularly in industries sensitive to energy consumption. The potential integration of AI-based systems for real-time monitoring and resource management could also be an area for further exploration, enhancing the adaptability and efficiency of hybrid computing architecture across various industrial applications.

REFERENCES

Abbaszadeh Shahri, A., Pashamohammadi, F., Asheghi, R., & Abbaszadeh Shahri, H. (2022). Automated Intelligent Hybrid Computing Schemes to Predict Blasting Induced Ground Vibration. *Engineering with Computers*, *38*, 3335–3349. https://doi.org/10.1007/s00366-021-01444-1

Agomuo, O. C., Jnr, O. W. B., & Muzamal, J. H. (2024). Energy-Aware AI-Based Optimal Cloud Infra Allocation for Provisioning of Resources. *27th IEEE/ACIS International Summer Conference on Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing, SNPD 2024 - Proceedings*, 269–274. https://doi.org/10.1109/snpd61259.2024.10673918

Ahmad, T., Zhu, H., Zhang, D., Tariq, R., Bassam, A., Ullah, F., AlGhamdi, A. S., & Alshamrani, S. S. (2022). Energetics Systems and Artificial Intelligence: Applications of Industry 4.0. *Energy Reports*, *8*, 334–361. https://doi.org/10.1016/j.egyr.2021.11.256

Alahakoon, D., Nawaratne, R., Xu, Y., De Silva, D., Sivarajah, U., & Gupta, B. (2023). Self-Building Artificial Intelligence and Machine Learning to Empower Big Data Analytics in Smart Cities. *Information Systems Frontiers*, *25*(1), 221–240. https://doi.org/10.1007/s10796-020-10056-x

Ali, O., Ishak, M. K., Bhatti, M. K. L., Khan, I., & Kim, K. Il. (2022). A Comprehensive Review of Internet of Things: Technology Stack, Middlewares, and Fog/Edge Computing Interface. *Sensors*, *22*(3), 995. https://doi.org/10.3390/s22030995

Anees, T., Habib, Q., Al-Shamayleh, A. S., Khalil, W., Obaidat, M. A., & Akhunzada, A. (2023). The Integration of WoT and Edge Computing: Issues and Challenges. *Sustainability (Switzerland)*, *15*(7), 5983. https://doi.org/10.3390/su15075983

Ayachi, R., Said, Y., & Ben Abdelali, A. (2021). Optimizing Neural Networks for Efficient FPGA Implementation: A Survey. *Archives of Computational Methods in Engineering*, *28*(7), 4537–4547. https://doi.org/10.1007/s11831-021-09530-9

Baccour, E., Mhaisen, N., Abdellatif, A. A., Erbad, A., Mohamed, A., Hamdi, M., & Guizani, M. (2022). Pervasive AI for IoT Applications: A Survey on Resource-Efficient Distributed Artificial Intelligence. *IEEE Communications Surveys and Tutorials*, *24*(4), 2366–2418. https://doi.org/10.1109/comst.2022.3200740

Bal, P. K., Mohapatra, S. K., Das, T. K., Srinivasan, K., & Hu, Y. C. (2022). A Joint Resource Allocation, Security with Efficient Task Scheduling in Cloud Computing Using Hybrid Machine Learning Techniques. *Sensors*, *22*(3), 1242. https://doi.org/10.3390/s22031242

Bobda, C., Mbongue, J. M., Chow, P., Ewais, M., Tarafdar, N., Vega, J. C., Eguro, K., Koch, D., Handagala, S., Leeser, M., Herbordt, M., Shahzad, H., Hofste, P., Ringlein, B., Szefer, J., Sanaullah, A., & Tessier, R. (2022). The Future of FPGA Acceleration in Datacenters and the Cloud. *ACM Transactions on Reconfigurable Technology and Systems*, *15*(3), 1–42. https://doi.org/10.1145/3506713

Devi, B. S. K., Vijayakumar, V., Suseela, G., Kavin, B. P., Sivaramakrishnan, S., & Rodrigues, J. J. P. C. (2022). An Improved Security Framework in Health Care Using Hybrid Computing. *Malaysian Journal of Computer Science*, *2022*(1), 50–61. https://doi.org/10.22452/mjcs.sp2022no1.4

Huang, Z., Fey, M., Liu, C., Beysel, E., Xu, X., & Brecher, C. (2023). Hybrid Learning-Based Digital Twin for Manufacturing Process: Modeling Framework and Implementation. *Robotics and Computer-Integrated Manufacturing*, *82*, 102545. https://doi.org/10.1016/j.rcim.2023.102545

Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial Intelligence Applications for Industry 4.0: A Literature-Based Study. *Journal of Industrial Integration and Management*, *7*(1), 83–111. https://doi.org/10.1142/S2424862221300040

Jiang, C., He, Z., Li, F., Xie, F., Zheng, L., Yang, J., & Yang, M. (2023). A Hybrid Computing Framework for Risk-Oriented Reliability Analysis in Dynamic PSA Context: A Case Study. *Quality and Reliability Engineering International*, *39*(8), 3445–3471. https://doi.org/10.1002/qre.3196

Katal, A., Dahiya, S., & Choudhury, T. (2023). Energy Efficiency in Cloud Computing Data Centers: A Survey on Software Technologies. *Cluster Computing*, *26*(3), 1845–1875. https://doi.org/10.1007/s10586-022-03713-0

Kaur, M., & Aron, R. (2022). An Energy-Efficient Load Balancing Approach for Scientific Workflows in Fog Computing. *Wireless Personal Communications*, *125*(4), 3549–3573. https://doi.org/10.1007/s11277-022-09724-9

Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022). Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions. *Journal of Network and Computer Applications*, *204*, 103405. https://doi.org/10.1016/j.jnca.2022.103405

Kong, X., Wu, Y., Wang, H., & Xia, F. (2022). Edge Computing for Internet of Everything: A Survey. *IEEE Internet of Things Journal*, *9*(23), 23472–23485. https://doi.org/10.1109/jiot.2022.3200431

Kristian, A., Sumarsan Goh, T., Ramadan, A., Erica, A., & Visiana Sihotang, S. (2024). Application of AI in Optimizing Energy and Resource Management: Effectiveness of Deep Learning Models. *International Transactions on Artificial Intelligence (ITALIC)*, *2*(2), 99–105. https://doi.org/10.33050/italic.v2i2.530

Li, J., Gu, C., Xiang, Y., & Li, F. (2022). Edge-Cloud Computing Systems for Smart Grid: State-of-the-Art, Architecture, and Applications. *Journal of Modern Power Systems and Clean Energy*, *10*(4), 805–817. https://doi.org/10.35833/mpce.2021.000161

Liu, B., Luo, Z., Chen, H., & Li, C. (2022). A Survey of State-of-the-Art on Edge Computing: Theoretical Models, Technologies, Directions, and Development Paths. *IEEE Access*, *10*, 54038–54063. https://doi.org/10.1109/access.2022.3176106

Memari, P., Mohammadi, S. S., Jolai, F., & Tavakkoli-Moghaddam, R. (2022). A Latency-Aware Task Scheduling Algorithm for Allocating Virtual Machines in a Cost-Effective and Time-Sensitive Fog-Cloud Architecture. *Journal of Supercomputing*, *78*(1), 93–122. https://doi.org/10.1007/s11227-021-03868-4

Mithas, S., Chen, Z. L., Saldanha, T. J. V., & De Oliveira Silveira, A. (2022). How Will Artificial Intelligence and Industry 4.0 Emerging Technologies Transform Operations Management? *Production and Operations Management*, *31*(12), 4475–4487. https://doi.org/10.1111/poms.13864

Muralidhar, R., Borovica-Gajic, R., & Buyya, R. (2022). Energy Efficient Computing Systems: Architectures, Abstractions and Modeling to Techniques and Standards. *ACM Computing Surveys*, *54*(11), 1–37. https://doi.org/10.1145/3511094

Murino, T., Monaco, R., Nielsen, P. S., Liu, X., Esposito, G., & Scognamiglio, C. (2023). Sustainable Energy Data Centres: A Holistic Conceptual Framework for Design and Operations. *Energies*, *16*(15), 5764. https://doi.org/10.3390/en16155764

Nain, G., Pattanaik, K. K., & Sharma, G. K. (2022). Towards Edge Computing in Intelligent Manufacturing: Past, Present and Future. *Journal of Manufacturing Systems*, *62*, 588–611. https://doi.org/10.1016/j.jmsy.2022.01.010

Raeisi-Varzaneh, M., Dakkak, O., Habbal, A., & Kim, B. S. (2023). Resource Scheduling in Edge Computing: Architecture, Taxonomy, Open Issues and Future Research Directions. *IEEE Access*, *11*, 25329–25350. https://doi.org/10.1109/access.2023.3256522

Rammer, C., Fernández, G. P., & Czarnitzki, D. (2022). Artificial Intelligence and Industrial Innovation: Evidence from German Firm-Level Data. *Research Policy*, *51*(7), 104555. https://doi.org/10.1016/j.respol.2022.104555

Rana, N., Jeribi, F., Khan, Z., Alrawagfeh, W., Ben Dhaou, I., Haseebuddin, M., & Uddin, M. (2024). A Systematic Literature Review on Contemporary And Future Trends in Virtual Machine Scheduling Techniques in Cloud and Multi-Access Computing. *Frontiers in Computer Science*, *6*, 1288552. https://doi.org/10.3389/fcomp.2024.1288552

Rodrigues Moreira, L. F., Moreira, R., Travençolo, B. A. N., & Backes, A. R. (2023). An Artificial Intelligence-as-a-Service Architecture for Deep Learning Model Embodiment on Low-Cost Devices: A Case Study of Covid-19 Diagnosis. *Applied Soft Computing*, *134*, 110014. https://doi.org/10.1016/j.asoc.2023.110014

Saiteja, P., & Ashok, B. (2022). Critical Review on Structural Architecture, Energy Control Strategies and Development Process Towards Optimal Energy Management in Hybrid Vehicles. *Renewable and Sustainable Energy Reviews*, *157*, 112038. https://doi.org/10.1016/j.rser.2021.112038

Saleem, M. U., Shakir, M., Usman, M. R., Bajwa, M. H. T., Shabbir, N., Shams Ghafaroki, P., & Daniel, K. (2023). Integrating Smart Energy Management System with Internet of Things and Cloud Computing for Efficient Demand Side Management in Smart Grids. *Energies*, *16*(12), 4835. https://doi.org/10.3390/en16124835

Solanki, P., Baldaniya, D., Jogani, D., Chaudhary, B., Shah, M., & Kshirsagar, A. (2022). Artificial Intelligence: New Age of Transformation in Petroleum Upstream. *Petroleum Research*, *7*(1), 106–114. https://doi.org/10.1016/j.ptlrs.2021.07.002

Stergiou, C. L., & Psannis, K. E. (2022). Digital Twin Intelligent System for Industrial IoT-Based Big Data Management and Analysis in Cloud. *Virtual Reality and Intelligent Hardware*, *4*(4), 279–291. https://doi.org/10.1016/j.vrih.2022.05.003

Surianarayanan, C., Lawrence, J. J., Chelliah, P. R., Prakash, E., & Hewage, C. (2023). A Survey on Optimization Techniques for Edge Artificial Intelligence (AI). *Sensors*, *23*(3), 1279. https://doi.org/10.3390/s23031279

Tatineni, S., & Chakilam, N. V. (2024). Integrating Artificial Intelligence with DevOps for Intelligent Infrastructure Management: Optimizing Resource Allocation and Performance in Cloud-Native Applications. *Journal of Bioinformatics and Artificial Intelligence*, *4*(1), 109–142. https://doi.org/10.1103/physrevlett.125.100501

Ullah, A., Kiss, T., Kovács, J., Tusa, F., Deslauriers, J., Dagdeviren, H., Arjun, R., & Hamzeh, H. (2023). Orchestration in the Cloud-to-Things Compute Continuum: Taxonomy, Survey and Future Directions. *Journal of Cloud Computing*, *12*(1), 135. https://doi.org/10.1186/s13677-023-00516-5

Zhang, H., Liu, J., Bai, J., Li, S., Luo, L., Wei, S., Wu, J., & Kang, W. (2022). HD-CIM: Hybrid-Device Computing-in-Memory Structure Based on MRAM and SRAM to Reduce Weight Loading Energy of Neural Networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *69*(11), 4465–4474. https://doi.org/10.1109/tcsi.2022.3199440