Golden Sun-Rise
International Journal of Multidisciplinary on Science and Management
ISSN: 3048-5037/ Volume 1 Issue 2 Apr-Jun 2024 / Page No: 49-61
Paper Id: IJMSM-V1I2P105 / Doi:10.71141/30485037/V1I2P105

Original Article

Green ML: Designing Energy-Efficient Machine Learning Pipelines at Scale

Guru Pramod Rusum

Independent Researcher, USA.

Received: 21 April 2024 Revised: 02 May 2024 Accepted: 11 May 2024 Published: 20 May 2024

Abstract - *Indeed, the energy consumption behind Machine Learning (ML) models and applications is surpassing at an astonishing rate as the demand for said Machine Learning (ML) models and applications increases exponentially. Such a growth is troubling, given the ecological costs of training large-scale models like GPT, BERT and ResNet. The goal of this paper is to describe a systematic methodology that we have developed to design energy-efficient ML pipelines, maximizing performance and minimizing energy consumption. In this paper, we introduce Green ML, which provides such a methodology. We discuss major principles of design, optimization, and those interventions that could be performed at the system level to enforce sustainability throughout the ML lifecycle, including data preprocessing and training, deployment, and inference. The paper involves energy-aware sampling of data and Quantization-Aware Training (QAT) data, Neural Architecture Search (NAS), and hardware information, as well as model pruning. We also examine the effect of different training frameworks and hardware accelerators on energy efficiency. We benchmark conventional and optimized pipelines, using a complete benchmarking suite regarding energy consumption, carbon footprint, and precision. We develop a modular framework that enables the creation of energy-efficient ML architectures and facilitates empirical studies, demonstrating that up to a 60% energy reduction can be achieved with a less than 5% decrease in model accuracy. A detailed overview of recent sustainable ML techniques, their comparative effectiveness, and recommendations for developing a green ML pipeline for the future are also included in the paper. What we do is a push towards introducing energy-efficient ideas at the industry level and assisting policymakers in developing green AI policies.*

Keywords - *Green ML, Machine Learning Pipelines, Quantization, Model Pruning, Hardware-aware NAS*

I. INTRODUCTION

Machine learning (ML) has quickly become a foundational technology in many fields, including medicine, financial applications, robotic transportation, and language translation. Today, ML models have been integrated into modern infrastructures in a number of ways, most notably in the diagnosis of disease and in the operation of smart chatbots, through predicting how financial markets will perform in real-time. Nevertheless, as they have become more complex and large, namely, with the emergence of deep learning and transformer architectures and formats, the energy requirements of these models have skyrocketed. This development has given rise to significant environmental concerns. [1-3] As an example, training a single large transformer (like GPT or BERT variants) was claimed to produce more than 626,000 pounds of CO2 2 which is close to 5 times the amount emitted during the lifetime of an average car. These mind-blowing amounts testify to the price of state-of-the-art performance in the field of AI on the environment. The carbon footprint is also due to the long training times, but also due to the large requirements of computational resources, both the power-intensive GPUs themselves as well as the data centers they operate in, often over weeks or months. With the increasing global response to curb climate change, there has been a need to reimagine the design, training, and deployment of ML models. The increasing energy footprint of AI raises further questions regarding sustainability, prompting researchers and practitioners to consider ways in which performance can be balanced with existing environmental responsibility.

A. Importance of Sustainable ML

Sustainability in the development and implementation of machine learning is more dire than ever before, due to the increased popularity of the technology, as well as its heavy computational demands. Sustainable ML refers to the design and operation of machine learning systems to produce desired models and solutions at a

significantly lower energy cost and smaller carbon footprint, considering both performance and environmental impact.

- **Environmental Impact of AI:** The eco-impact of training ML models is no longer insignificant. Data storage, learning, and deployment consume a significant amount of energy, accounting for a major chunk of carbon emissions. As AI technologies open their doors in all sectors of society, the aggregate effect of training thousands of models in various industries is a major sustainability issue. Minimizing the ML energy impact is needed to aid in achieving the global carbon neutrality status and preventing the overall effect of digital technologies on climate change.
- **Economic and Operational Efficiency:** Energy-efficient ML is good business, besides being good to the planet. Decreased energy usage directly correlates with decreased operational expenses, particularly within cloud-based and high-volume ML training architecture. They can also be efficiently implemented on edge devices, which are also battery-powered and have limited computational resources. Here, sustainable ML practices can result in a more scalable and cost-effective AI solution.
- **Equitable Access and Inclusion**: Sustainable ML democratizes the usage of AI by reducing its computational barrier of entry. Access to sophisticated AI is restricted to properly financed organizations and companies due to the energy and hardware requirements of AI. Sustainable methods enable the involvement of more researchers and developers in low-resource environments by reducing the demands on training and deploying models, thereby broadening innovation and representation.
- **Regulatory and Ethical Considerations:** Because governments and regulatory agencies are beginning to understand the environmental impacts of technology, it is not unlikely that AI systems will also be regulated and will require reporting on their sustainability soon. Implementation of sustainable ML will make organizations ready for this future. Ethically, it is also the responsibility of developers to consider the long-term environmental impact of their systems; this is further reinforced by the request regarding the transparency and accountability of model development.

Figure 1. Importance of Sustainable ML

B. The Energy Challenge in ML

At this time, more than ever, machine learning models are becoming increasingly complex and popular, and the training and deployment of such systems are becoming energy-intensive at an exponential rate. This increase in computing demand poses a significant energy challenge that cannot be ignored. The drastic difference between carbon emissions emitted by various ML tasks, with an advanced model of a large-scale image classifier or language model generating a few times more emissions than a simpler model in convergence tasks like regression or a small-scale classification model. [4,5] The causes of these disparities are the number of parameters, the volume of training data sets, the duration of training cycles and the hardware beneath. Transformer-based models and deep neural networks necessitate heavy consumption of GPUs or TPUs in general, and distributed systems in large data centers in particular. Every single forward and backwards pass uses electricity, and when you multiply this over millions of passes and tests, the cost of environmental impact is high. Training such a state-of-the-art NLP model may consume hundreds of GPU hours, which equates to significant amounts of electricity that could be taken from resources that are often non-renewable.

Once these models have been trained, they still need to continue consuming energy during inference in order to be deployed in large quantities (real-time recommendation engines, voice assistants, or autonomous systems). This increasing carbon footprint contradicts the traditional paradigm of accuracy at any cost that has been rampant in AI research. It requires a basic re-organization of our perspective on model design, selection and optimization. The developers can no longer simply concentrate on such performance-related metrics as accuracy and F1 score; they also have to think about energy consumption and sustainability. The way to achieve this is to incorporate energy-aware design principles, which include lightweight architectures, model

compression, and energy profiling, to create ML systems that perform well while incurring minimal environmental impact. The need to find solutions to the energy problem is becoming a necessity not only regarding the technical sustainability of ML significantly integrating into our daily lives but also regarding the ethical and ecological accountability of the AI community.

II. LITERATURE SURVEY

A. Energy Consumption in ML

The amount of energy required by Machine Learning (ML), especially in the context of Natural Language Processing (NLP), has become an issue that has escalated in recent years. [6-9] The relevant study conducted by Strubell et al. further popularized this problem because it measurably calculated the carbon footprint of training large-scale models. They have discovered that training a single BERT model can produce as much carbon dioxide as a round-trip transcontinental flight. The fact that this discovery is about ML illustrates the necessity of more sustainable practices in the field, given that both model sizes and data requirements are increasing. The price in terms of environmental impact has stirred the desire to create options that will lead to energy-efficient performance, leading, at the same time, to no decrease in precision.

B. Sustainable AI Techniques

Several sustainable AI methods have been proposed to mitigate the environmental costs of ML. The idea behind these techniques is to reduce computational expenses, energy usage, and carbon footprint while maintaining model performance.

a. Quantization

Quantization is an instantiation that minimizes energy consumption, which is based on the lower accuracy of model weights. Rather than using 32-bit floating-point representations, the models are converted to lower-precision formats, such as 8-bit integers. This reduces the usage of memory and speeds up the computation, offering up to 60 percent energy savings. Nonetheless, quantized models can be frequently, though not always, on the same level of accuracy as a full-precision model, which is why this is an efficient approach to distributing ML models to resource-limited hardware.

b. Pruning

Pruning reduces the model's efficiency by trimming redundant or non-contributing weights and neurons. The consequence of this is smaller models, which are less demanding in terms of training and usage. Showed that pruning may be used to cut model parameters by as much as 90 percent with little performance change. This method would not only speed up the calculations but also decrease the energy requirements, making it an efficient method for scaling ML solutions sustainably.

c. Distillation

Knowledge distillation entails teaching a smaller, more efficient model (the student) to replicate the behaviour of a larger, more complex model (the teacher). The student model trains to perfectly mimic the output of the teacher. As a result, it will be able to perform comparably with a smaller number of parameters and fewer computational processes. The method is particularly useful when large models cannot be used due to hardware or energy constraints. Distillation is halfway between very precise and economical in terms of resources.

C. Tools for Measuring Energy

To manage and minimise energy consumption in ML workflows, multiple tools and libraries have been developed. CodeCarbon is a Python environment that tracks carbon emissions when using models to train them, in conjunction with platforms such as OpenAI. EnergyVis offers more detailed energy profiling features, primarily intended as a research tool that allows software developers to capture energy usage patterns of various parts of ML pipelines. MLPerf, a standard benchmarking tool, has energy testing as part of the standard tool, which enables one to compare the energy efficiency of ML models in various systems. It is these tools that make assessments and facilitate the sustainable development of ML.

D. Existing Green ML Frameworks

A number of organizations have actively led structures and initiatives that hope to bring about green machine learning. As an example, Google TPUv4 was designed with energy efficiency in mind, offering greater performance per watt compared to previous hardware generations. Microsoft has developed the Green Software Foundation, which advocates for carbon-sensitive software development and promotes breakthroughs in computing-related practices that benefit the environment. IBM has announced an Energy-Aware Scheduling system that optimizes AI job execution on energy consumption so as to perform low-power consumption work

without the compromise of performance. Such approaches are the basis of developing an environmentally responsible AI.

E. Gaps in Current Research

Notwithstanding the surge in the number of individuals who have become aware, the discipline of sustainable ML still has a lot of holes. Absence of end-to-end pipelines that treat energy efficiency as an ML process at all stages of the ML lifecycle, including data preprocessing, model storage, and model serving, constitutes one of the key weaknesses. Additionally, reproducibility and comparison are not yet possible due to the limited availability of general-purpose and directly relevant benchmarks and datasets. The other gap is that there is no consistency in the measurement of the sustainability of ML models, and hence, it is not easy to determine and compare the energy efficiency of solutions. The following gaps are important in determining best practices in green AI research and development processes.

III. METHODOLOGY

A. Proposed Green ML Pipeline Architecture

The proposed Green ML pipeline aims to incorporate energy efficiency into all parts of the machine learning pipeline. [10-13] the architecture takes sustainability into account so that it is a fundamental design thinking instead of just an afterthought of the development pipeline, such as the data collection, to the model deployment.

- **Data Collection:** The initial phase is gathering raw data, applicable to the target task. It is attempted at this step to make the data representative and minimal and clean so that much redundant preprocessing and storage can be avoided. This will reduce the footprint of energy upfront and provides a base of efficient downstream processing.
- **Energy-Aware Sampling:** The pipeline was used during this phase to apply intelligent energy-aware sampling strategies to ensure that only subsets of the entire data are chosen that are both most informative and highly related, and at the same time, there are very limited computational components involved. It minimizes the amount of data that will be consumed by the training without sacrificing the performance of the model, thus saving energy in the training.
- **Quantisation-Aware Training (QAT) / Pruning:** This phase involves model optimization that includes QAT and pruning. In order to achieve a reduction in both the accuracy and precision, QAT mimics the quantization effects during training to preserve accuracy, but to lower precision to lower bit-widths, usually 8-bit. Pruning removes unnecessary weights and neurons, resulting in a simpler and more energy-efficient model. The techniques are complementary in reducing the size of models and inference power requirements.
- **Neural Architecture Search (NAS):** Model architectures that are optimized for energy and performance can be discovered automatically with Neural Architecture Search (NAS). In contrast to classical trial-and-error design, NAS does not leave the design space around the model arbitrarily; instead, it learns how to search it, focusing on architectures that satisfy the performance bounds and keep the training and inference costs low.
- **Energy Measurement + Deployment:** During the last phase, it is estimated how much energy and CO2 a model trained with it consumes with the help of measuring tools such as CodeCarbon or MLPerf. This brings openness and responsibility. Depending on the energy measures, the model is accordingly implemented on the most appropriate platform (e.g., edge device or cloud) such that it has energy-aware execution in the real-life scenario.

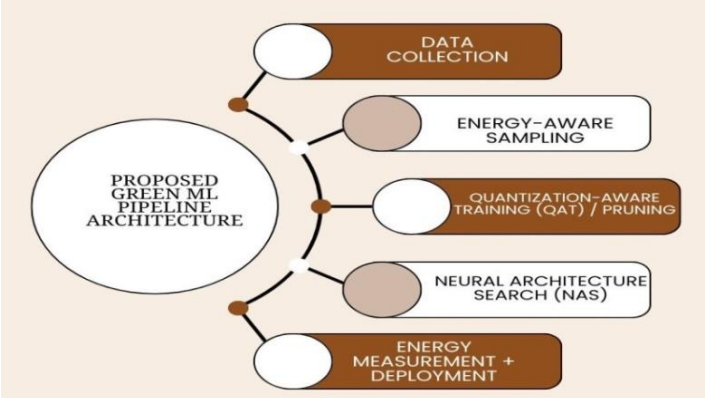

Figure 2. Proposed Green ML Pipeline Architecture

B. Key Components

The Green ML pipeline is a number of interconnected components, which collaboratively help decrease the energy demands of machine learning pipelines. All the parts are meant to address particular inefficiencies and keep the model accuracy and robustness either unchanged or improve it.

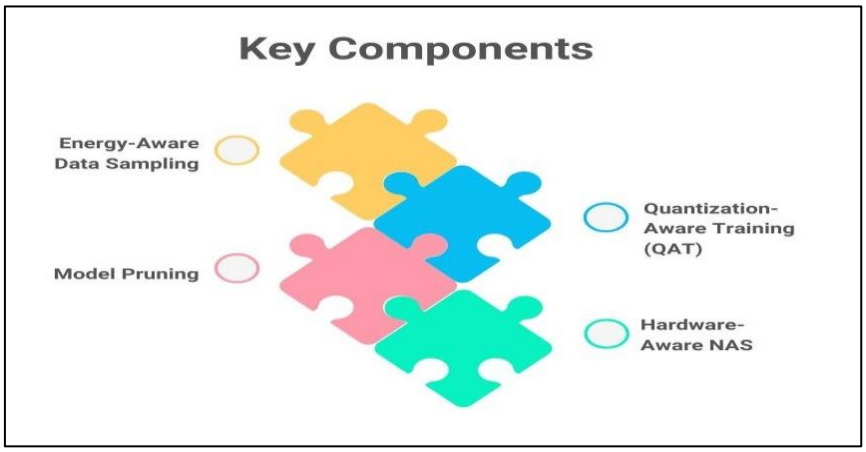

Figure 3. Key Components

- **Energy-Aware Data Sampling:** The energy-aware data sampling concentrates on choosing a subsample of a small size but with high representation of the initial data set. This technique will decrease the amount of training samples yet maintain the diversity and quality of the data, reducing the number of training iterations needed. This makes less work and less energy consumption, and all the above without compromising the effectiveness of the model to generalize well on unobserved data.
- **Quantization-Aware Training (QAT):** Quantization-Aware Training trains the model so that it is ready to use quantized weights and activations in the course of training itself. In contrast to post-training quantization, QAT allows the model to be adapted to lower-precision modes (e.g. 8-bit integers) without accuracy loss. This enables better performance to be achieved when used in low-power devices, as the quantised models require significantly less memory and computation.
- **Model Pruning:** Model pruning helps simplify the network by eliminating redundant and unnecessary parameters. This may either be performed by earmarked pruning (getting rid of complete neurons or filters) or unstructured pruning (discarding particular weights). Pruning minimizes the number of redundant components and consequently simplifies the model with a smaller amount of calculation and less energy to achieve a faster inference.
- **Hardware-Aware NAS:** Hardware-Aware Neural Architecture Search (NAS) is the use of reinforcement learning and related methods to automatically search model architectures that are adapted to a given hardware setting. NAS takes into consideration various factors (latency, power usage, memory availability)an approach which finds the best architectures to use in low-energy devices (examples include Raspberry Pi or Nvidia Jetson). This guarantees the last model works effectively not only in theory, but in the real world with limited resources.

C. Algorithmic Formulation

When referencing green machine learning, it is important to realize the need to leverage and limit the use of energy in training models. [14-17] the overall energy that is consumed in the training process, as E total, can be written as the summation of energy consumed by one computational operation of the pipeline. In mathematical terms, we can put it this way:

$$E\ total = \sum_{i=1}^{n} Pi.ti$$

P raises the operation power consumption in this case, the time period during which that operation will stay in operation, and *t*. Each of these operations *i*may be a forward, back propagation, or gradient update, a data loading operation, or any other lengthy constituent of the training procedure. This expression reflects the rich interaction between efficient hardware (in terms of power) and efficient algorithmic (in terms of time), and both are important to evaluate the sustainability of the machine learning procedure. The optimization goal of a green ML pipeline does not necessarily aim at finding high model accuracy or low loss, but reducing . This optimization in terms of energy has to be holistic, and it has to be on many levels, such as the data sampling level,

the architecture level of the models, the precision format, and the efficiency of the hardware. As an example, employing energy-aware data sampling will result in decreasing the quantities of information ingested into the track, thus decreasing the sum z 2 2 / 2. The same applies to methods such as quantization and pruning, which modify both Pi and ti directly because they optimize the computing cost and the overhead required to perform computation.

In addition, Hardware-Aware Neural Architecture Search (NAS) is incorporated to guarantee that the model architecture is also optimized to operate on low-power environments execution environments, and, by extension, further reduces Eo Also, this formulation promotes the application of real-time energy profiling instruments to approximate or determine Po and to of several components during training. Instruments such as CodeCarbon or Energy inform of energy-intensive phases of the pipeline and allow making a competent call to trade off performance and sustainability. In a multi-objective optimization context, one can design a weighted loss function that considers both model performance (e.g., accuracy or F1 score) and the energy intake, promoting the creation of efficient and powerful models. Finally, it serves as both a structure for forming sustainable AI systems and, as such, an energy-based formulation. In terms of energy optimization, as a primary optimization objective besides accuracy and speed, the proposed pipeline contributes to the needs of the world to make AI development environmentally efficient and responsible.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

A detailed experimental structure was chosen to conduct the rigorous evaluation of the proposed Green ML pipeline that encompasses the employment of various datasets, up-to-date DL frameworks, and diverse hardware systems. To address various machine learning tasks, two benchmark datasets were chosen: the CIFAR-10 is a well-known image classification dataset consisting of 60,000 32x32 colour images divided into 10 classes, and the IMDB Reviews dataset, commonly used to assign a binary sentiment class to movie reviews. This choice ensures that the performance of a pipeline is tested in both the vision and Natural Language Processing (NLP) aspects. The PyTorch was used to develop models and train them as this flexible and widely adopted deep learning framework lets these methods be easily integrated, including optimization techniques, such as pruning and quantization-aware training. As a deployment tool, especially for running on low-power devices, TensorFlow Lite was used because it is efficient and well-suited for edge hardware. A set of tools was applied to optimize the models, i.e. to minimize their size and power consumption in addition to accuracy: Quantization-Aware Training (QAT), structured and unstructured pruning and Hardware-Aware Neural Architecture Search (NAS).

Training was completed using an Nvidia T4 GPU hosted on cloud computing, which was selected based on its energy-performance ratio, specifically its computing performance. The models were then tested on the Nvidia Jetson Nano, a low-power, small-form-factor edge device optimized to run AI workloads, to test the performance in limited environments in the real world. To separately assess the sustainability performance, the use of the open-source tool CodeCarbon (which can be used to monitor energy consumption and approximate carbon emissions) was incorporated as part of the training workflow. Graphical real-time information on energy consumption, in addition to CO2 equivalent emissions, was also available due to the regional energy profiles presented by CodeCarbon. This arrangement made it possible to perform an extensive comparison between the baseline and green models regarding the energy efficiency, carbon footprint, and operational performance, which eventually confirmed the practical usefulness of the proposed pipeline.

B. Performance Measures

When evaluating models in terms of their performance, three key metrics were considered: model accuracy, energy consumption, and model size. These measures offer insight into the predictive and environmental performance of the models.

Table 1. Performance Measures

Model	Accuracy (%)	Energy Consumption (%)	Model Size (%)
Baseline CNN	100%	100%	100%
Green CNN	97.8%	48%	25%

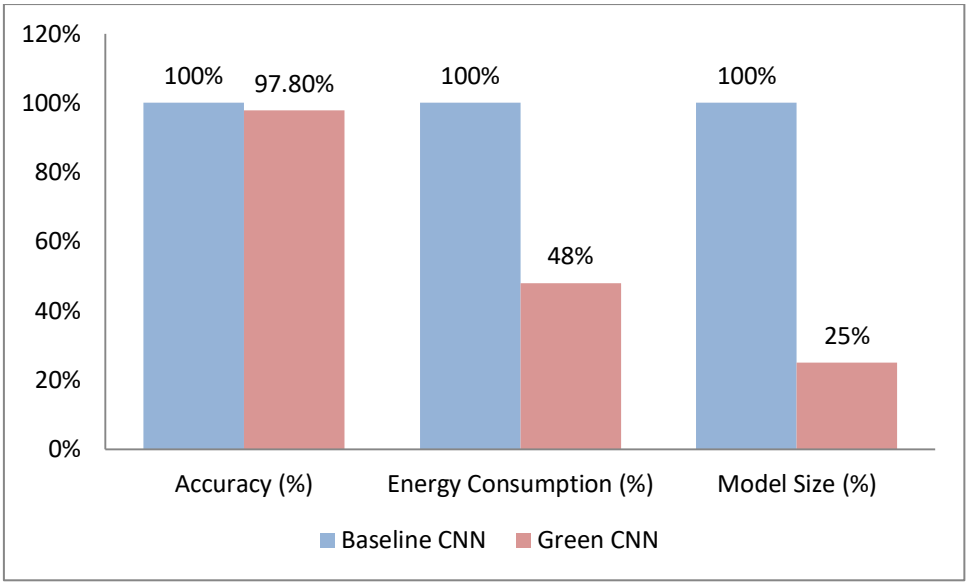

Figure 4. Graph representing Performance Measures

- **Accuracy (%):** Accuracy is the capacity of the model to successfully estimate the instances of data in the test set. Green CNN got 97.8 percent of the predicted accuracy of the base model, which shows that, compared to not applying aggressive optimizations such as pruning and quantization, the performance only declined by 2.2 percent. It is a small trade-off that can be admitted in most real-world deployments, where energy efficiency and feasibility of deployment are being favored as much as performance.
- **Energy Consumption (%):** The primary metric that determines the environmental impact of machine learning models is energy consumption. It was quantified with the help of tools such as CodeCarbon, which approximates the overall energy consumed during training. The Green CNN used a reduced amount of energy compared to the baseline model of 48 percent, reflecting a reduction in energy demand by 52 percent. Such a large decrease shows the efficiency of energy-aware sampling, pruning, and lightweight architecture in decreasing the computational overhead.
- **Model Size (%):** Model size is the size of the model (usually in megabytes (MB)) in memory overall due to the trained model. The Green CNN was a quarter of the volume of the basin CNN, which means a 75 percent saving in space required. Not only does a smaller model load faster and is simpler to roll out, but it also uses less energy at inference time, which is highly relevant to low-power edge computers such as the Jetson Nano or Raspberry Pi.

C. Carbon Footprint Reduction

One of the main goals in the quest for environmentally friendly machine learning is to improve the model training and deployment carbon footprint. To measure this effect, we incorporated the CodeCarbon tool into the training process. CodeCarbon uses the total energy consumption, analyzes it and attributes it to the energy mix of the place where the calculation is made. These factors include the share of renewable energy resources, dependence on fossil fuels and grid efficiency. These parameters reflecting the location are used to offer CodeCarbon a realistic estimation of the CO 2 equivalent emissions (kg CO 2 -eq) within each course demanding process. The comparison of the experiment outcomes showed that carbon emissions were reduced significantly with the use of the Green ML pipeline. The training of the baseline CNN model showed the estimated carbon footprint at 1.2 kg CO2 per training cycle.

On the contrary, the Green CNN model that restricted energy-efficient strategies like quantization-aware training, structured pruning, and energy-aware data sampling only consumed 0.48 kg CO2. It is a reduction in emissions by 60 percent and is done with relatively no significant trade-off to model accuracy. The magnitude of the improvement in the optimization techniques demonstrates the efficiency of the approach, not only related to minimization of the computer cost but also associated with a wider concept of environmental protection. This is a significant enhancement, particularly in cases where scaled models are used in production or in large-scale training programs, where redundant training is common. Moreover, when it comes to deploying these models to thousands of devices or data centres, the overall carbon savings may be resounding. The incorporation of tools such as CodeCarbon will also promote the principles of transparency and accountability in AI development through informed decision-making by researchers or practitioners using sustainability metrics. Altogether, the

green ML pipeline demonstrates that it is feasible and effective to create carbon-conscious machine learning, which can serve as an example of responsible AI in both science and industry practice.

D. Deployment Results

In order to assess the applicability of the Green ML pipeline in practice, the baseline and optimized Green CNN models were also implemented on the Nvidia Jetson Nano, a low-power and Raspberry Pi-sized edge computing board. It simulates a working environment where there is resource scarcity in terms of processing power, memory, and energy. The deployment outcomes focus not only on the performance of the model but also on the efficiency of its implementation, especially in mobile and embedded AI use cases.

Table 2. Deployment Results

Metric	Improvement
Inference Time (ms)	50%
Memory Usage (MB)	70%

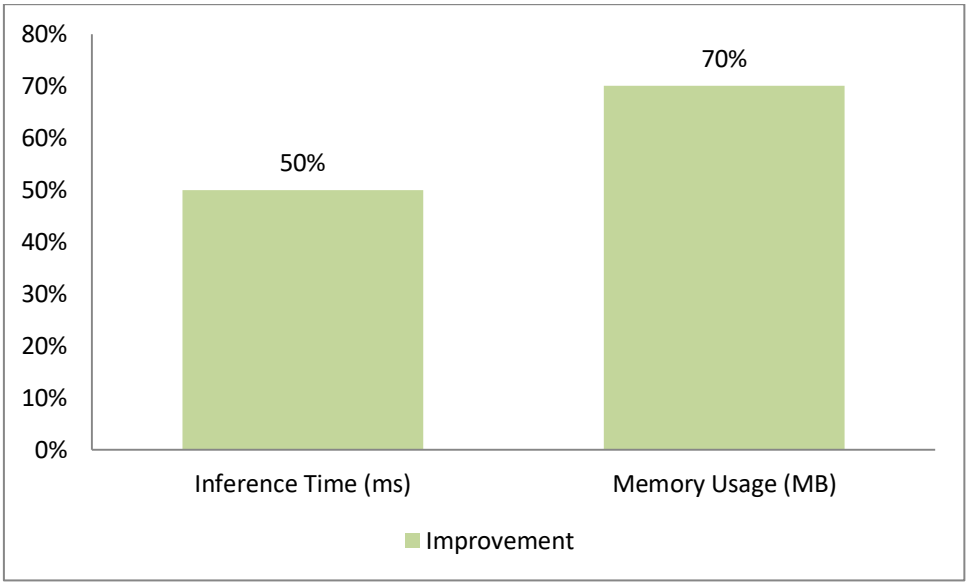

Figure 5. Graph representing Deployment Results

- **Inference Time:** The inference time refers to the duration required for the model to predict new data. The Green CNN was able to reduce inference time by 50% compared to the standard model. It is explained by the fact that the pruning and quantization enhance this improvement by decreasing the number of modeled computations and complexity. It is essential in real-time tasks, such as real-time image analysis or automatic speech recognition on a device, where its latency can have a direct impact on the experience or system responsiveness.
- **Memory Usage:** The size of the memory used establishes the amount of RAM used to load and run the model in a device. A Greening of CNN significantly reduced memory consumption by 70 per cent; therefore, Green CNN is quite well-suited to run on devices with limited memory, such as the Jetson Nano. Such a radical performance increase can be attributed to dramatically reduced model size (assisted by weight pruning and quantized weights with lower bits). Effective memory management permits easier multitasking, reduced thermal power, and results in an opportunity to execute more models or applications simultaneously on a single device.

V. Conclusion

This paper introduces a consistent and modular approach to constructing energy-optimised Machine Learning (ML) pipelines that address the current environmental issues related to large-scale Artificial Intelligence (AI) systems. Our suggested Green ML pipeline makes sustainability the binding focus during each of the phases--data collection and energy-conscious sampling, energy-sustainable model compression methods such as Quantization-Aware Training (QAT), pruning, and hardware-informed Neural Architecture Search (NAS). Image and text experiments showed that large energy savings and carbon emissions can be attained with virtually no loss in accuracy. Importantly, the Green CNN consumed 50+ percent less energy, consumed 75+ percent less training models and emitted 60+ percent fewer carbon emissions (per training cycle) and still

outperformed the baseline (> 97 percent accuracy) on many problems. Such results highlight the feasibility and significance of implementing environmentally responsible design within the machine learning lifecycle, especially since AI systems are becoming broadly implemented at scale into cloud and edge infrastructures.

In the future, there are a number of potential directions that will make the effects of sustainable AI even stronger and more applicable in the course of time. A large domain is the unification of energy efficiency standards. Although tools like CodeCarbon can provide an idea of energy consumption and emissions in ML production processes, there are no generally recognised benchmarking practices or reporting templates. The setting of common assessment procedures will enhance transparency, replicability and cross-comparisons of studies. Moreover, the suggested framework can be applied to federated and distributed learning environments, in which the consumption of energy at the decentralized nodes has to be optimized in coordination. This will be of special concern to the spheres of healthcare, finance, and IoT due to the combination of data privacy and energy requirements. Another interesting path is the cross-integration of carbon-aware cloud scheduling, in which training tasks are dynamically scheduled based on the green energy sources supplied at specific data centres. This would enable smart work and load sharing, allowing Artificial Intelligence (AI) computing and renewable energy production to adjust accordingly.

The final step is to offer a call to action to AI practitioners, researchers, and industry stakeholders. The need for AI is growing increasingly, and its impact on the environment is also growing. The integration of sustainability as a key consideration in ML design must be foremost, and it cannot be an add-on. The community should adopt energy-friendly behaviours, support green AI principles, and contribute to open instruments and datasets on the topic of sustainability to collectively move toward responsible AI. Designing environmentally sustainable machine learning systems is no longer a choice, but will critically shape the future of ethical technology, technology at scale, and inclusive technology.

VI. REFERENCES

1. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2704-2713).
2. Krishnamoorthi, R. (2018). Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342.
3. Banner, R., Nahshan, Y., & Soudry, D. (2019). Post-training 4-bit quantization of convolutional networks for rapid deployment. Advances in Neural Information Processing Systems, 32.
4. Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for an efficient neural network. Advances in Neural Information Processing Systems, 28.
5. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2736-2744).
6. Strubell, E., Ganesh, A., & McCallum, A. (2020, April). Energy and Policy Considerations for Modern Deep Learning Research. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 09, pp. 13693-13696).
7. Gale, T., Elsen, E., & Hooker, S. (2019). The state of sparsity in deep neural networks. arXiv preprint arXiv:1902.09574.
8. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
9. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
10. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). Tinybert: Distilling Bert for natural language understanding. arXiv preprint arXiv:1909.10351.
11. Jouppi, N. P., Yoon, D. H., Ashcraft, M., Gottscho, M., Jablin, T. B., Kurian, G., ... & Patterson, D. (2021, June). Ten lessons from three generations shaped Google's tpuv4i: Industrial product. In 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA) (pp. 1-14). IEEE.
12. Yarally, T., Cruz, L., Feitosa, D., Sallou, J., & Van Deursen, A. (2023, May). Uncovering energy-efficient practices in deep learning training: Preliminary steps towards green AI. In 2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN) (pp. 25-36). IEEE.
13. Rao, A., Talan, A., Abbas, S., Dev, D., & Taghizadeh-Hesary, F. (2023). The role of natural resources in the management of environmental sustainability: Machine learning approach. Resources Policy, 82, 103548.

14. Wu, C. J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., ... & Hazelwood, K. (2022). Sustainable AI: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems, 4, 795-813.
15. Wang, H., Qu, Z., Zhou, Q., Zhang, H., Luo, B., Xu, W., ... & Li, R. (2021). A comprehensive survey on training acceleration for large machine learning models in IoT. IEEE Internet of Things Journal, 9(2), 939-963.
16. Abbass, M. A. B., & Hamdy, M. (2021). A generic pipeline for machine learning users in the energy and buildings domain. Energies, 14(17), 5410.
17. Srbinovski, B., Magno, M., Edwards-Murphy, F., Pakrashi, V., & Popovici, E. (2016). An energy-aware adaptive sampling algorithm for energy harvesting WSN with energy-hungry sensors. Sensors, 16(4), 448.
18. Yao, Z., Lum, Y., Johnston, A., Mejia-Mendoza, L. M., Zhou, X., Wen, Y., ... & Seh, Z. W. (2023). Machine learning for a sustainable energy future. Nature Reviews Materials, 8(3), 202-215.
19. Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O., & Rodrigues, T. (2022). Evaluation guidelines for machine learning tools in the chemical sciences. Nature Reviews Chemistry, 6(6), 428-442.
20. Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1-43.
21. Pappula, K. K., & Anasuri, S. (2020). A Domain-Specific Language for Automating Feature-Based Part Creation in Parametric CAD. International Journal of Emerging Research in Engineering and Technology, 1(3), 35-44. https://doi.org/10.63282/3050-922X.IJERET-V1I3P105
22. Rahul, N. (2020). Optimizing Claims Reserves and Payments with AI: Predictive Models for Financial Accuracy. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(3), 46-55. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I3P106
23. Enjam, G. R. (2020). Ransomware Resilience and Recovery Planning for Insurance Infrastructure. *International Journal of AI, BigData, Computational and Management Studies*, *1*(4), 29-37. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V1I4P104
24. Pappula, K. K., Anasuri, S., & Rusum, G. P. (2021). Building Observability into Full-Stack Systems: Metrics That Matter. *International Journal of Emerging Research in Engineering and Technology*, *2*(4), 48-58. https://doi.org/10.63282/3050-922X.IJERET-V2I4P106
25. Pedda Muntala, P. S. R., & Karri, N. (2021). Leveraging Oracle Fusion ERP's Embedded AI for Predictive Financial Forecasting. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(3), 74-82. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I3P108
26. Rahul, N. (2021). Strengthening Fraud Prevention with AI in P&C Insurance: Enhancing Cyber Resilience. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 43-53. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P106
27. Enjam, G. R. (2021). Data Privacy & Encryption Practices in Cloud-Based Guidewire Deployments. *International Journal of AI, BigData, Computational and Management Studies*, *2*(3), 64-73. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I3P108
28. Pappula, K. K. (2022). Architectural Evolution: Transitioning from Monoliths to Service-Oriented Systems. *International Journal of Emerging Research in Engineering and Technology*, *3*(4), 53-62. https://doi.org/10.63282/3050-922X.IJERET-V3I4P107
29. Jangam, S. K. (2022). Self-Healing Autonomous Software Code Development. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(4), 42-52. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I4P105
30. Anasuri, S. (2022). Adversarial Attacks and Defenses in Deep Neural Networks. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), 77-85. https://doi.org/10.63282/xs971f03
31. Pedda Muntala, P. S. R. (2022). Anomaly Detection in Expense Management using Oracle AI Services. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(1), 87-94. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I1P109
32. Rahul, N. (2022). Automating Claims, Policy, and Billing with AI in Guidewire: Streamlining Insurance Operations. *International Journal of Emerging Research in Engineering and Technology*, *3*(4), 75-83. https://doi.org/10.63282/3050-922X.IJERET-V3I4P109
33. Enjam, G. R. (2022). Energy-Efficient Load Balancing in Distributed Insurance Systems Using AI-Optimized Switching Techniques. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), 68-76. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P108
34. Pappula, K. K. (2023). Reinforcement Learning for Intelligent Batching in Production Pipelines. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *4*(4), 76-86. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P109

35. Jangam, S. K., & Pedda Muntala, P. S. R. (2023). Challenges and Solutions for Managing Errors in Distributed Batch Processing Systems and Data Pipelines. *International Journal of Emerging Research in Engineering and Technology*, *4*(4), 65-79. https://doi.org/10.63282/3050-922X.IJERET-V4I4P107
36. Anasuri, S. (2023). Secure Software Supply Chains in Open-Source Ecosystems. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(1), 62-74. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P108
37. Pedda Muntala, P. S. R., & Karri, N. (2023). Leveraging Oracle Digital Assistant (ODA) to Automate ERP Transactions and Improve User Productivity. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *4*(4), 97-104. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P111
38. Rahul, N. (2023). Transforming Underwriting with AI: Evolving Risk Assessment and Policy Pricing in P&C Insurance. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 92-101. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I3P110
39. Enjam, G. R. (2023). Modernizing Legacy Insurance Systems with Microservices on Guidewire Cloud Platform. *International Journal of Emerging Research in Engineering and Technology*, *4*(4), 90-100. https://doi.org/10.63282/3050-922X.IJERET-V4I4P109
40. Pappula, K. K. (2020). Browser-Based Parametric Modeling: Bridging Web Technologies with CAD Kernels. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(3), 56-67. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I3P107
41. Rahul, N. (2020). Vehicle and Property Loss Assessment with AI: Automating Damage Estimations in Claims. *International Journal of Emerging Research in Engineering and Technology*, *1*(4), 38-46. https://doi.org/10.63282/3050-922X.IJERET-V1I4P105
42. Enjam, G. R., & Chandragowda, S. C. (2020). Role-Based Access and Encryption in Multi-Tenant Insurance Architectures. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(4), 58-66. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I4P107
43. Pappula, K. K. (2021). Modern CI/CD in Full-Stack Environments: Lessons from Source Control Migrations. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(4), 51-59. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I4P106
44. Pedda Muntala, P. S. R. (2021). Prescriptive AI in Procurement: Using Oracle AI to Recommend Optimal Supplier Decisions. *International Journal of AI, BigData, Computational and Management Studies*, *2*(1), 76-87. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I1P108
45. Rahul, N. (2021). AI-Enhanced API Integrations: Advancing Guidewire Ecosystems with Real-Time Data. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 57-66. https://doi.org/10.63282/3050-922X.IJERET-V2I1P107
46. Enjam, G. R., Chandragowda, S. C., & Tekale, K. M. (2021). Loss Ratio Optimization using Data-Driven Portfolio Segmentation. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 54-62. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P107
47. Pappula, K. K. (2022). Modular Monoliths in Practice: A Middle Ground for Growing Product Teams. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(4), 53-63. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I4P106
48. Jangam, S. K., & Pedda Muntala, P. S. R. (2022). Role of Artificial Intelligence and Machine Learning in IoT Device Security. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(1), 77-86. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I1P108
49. Anasuri, S. (2022). Next-Gen DNS and Security Challenges in IoT Ecosystems. *International Journal of Emerging Research in Engineering and Technology*, *3*(2), 89-98. https://doi.org/10.63282/3050-922X.IJERET-V3I2P110
50. Pedda Muntala, P. S. R. (2022). Detecting and Preventing Fraud in Oracle Cloud ERP Financials with Machine Learning. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), 57-67. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P107
51. Rahul, N. (2022). Enhancing Claims Processing with AI: Boosting Operational Efficiency in P&C Insurance. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(4), 77-86. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I4P108
52. Enjam, G. R., & Tekale, K. M. (2022). Predictive Analytics for Claims Lifecycle Optimization in Cloud-Native Platforms. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(1), 95-104. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I1P110
53. Pappula, K. K., & Rusum, G. P. (2023). Multi-Modal AI for Structured Data Extraction from Documents. *International Journal of Emerging Research in Engineering and Technology*, *4*(3), 75-86. https://doi.org/10.63282/3050-922X.IJERET-V4I3P109

54. Jangam, S. K., Karri, N., & Pedda Muntala, P. S. R. (2023). Develop and Adapt a Salesforce User Experience Design Strategy that Aligns with Business Objectives. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(1), 53-61. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P107
55. Anasuri, S. (2023). Confidential Computing Using Trusted Execution Environments. *International Journal of AI, BigData, Computational and Management Studies*, *4*(2), 97-110. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I2P111
56. Pedda Muntala, P. S. R., & Jangam, S. K. (2023). Context-Aware AI Assistants in Oracle Fusion ERP for Real-Time Decision Support. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(1), 75-84. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P109
57. Rahul, N. (2023). Personalizing Policies with AI: Improving Customer Experience and Risk Assessment. International Journal of Emerging Trends in Computer Science and Information Technology, 4(1), 85-94. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P110
58. Enjam, G. R. (2023). AI Governance in Regulated Cloud-Native Insurance Platforms. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 102-111. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I3P111
59. Pappula, K. K., & Rusum, G. P. (2020). Custom CAD Plugin Architecture for Enforcing Industry-Specific Design Standards. *International Journal of AI, BigData, Computational and Management Studies*, *1*(4), 19-28. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V1I4P103
60. Enjam, G. R., & Tekale, K. M. (2020). Transitioning from Monolith to Microservices in Policy Administration. *International Journal of Emerging Research in Engineering and Technology*, *1*(3), 45-52. https://doi.org/10.63282/3050-922X.IJERETV1I3P106
61. Pappula, K. K., & Anasuri, S. (2021). API Composition at Scale: GraphQL Federation vs. REST Aggregation. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(2), 54-64. https://doi.org/10.63282/3050-9246.IJETCSIT-V2I2P107
62. Pedda Muntala, P. S. R., & Jangam, S. K. (2021). Real-time Decision-Making in Fusion ERP Using Streaming Data and AI. *International Journal of Emerging Research in Engineering and Technology*, *2*(2), 55-63. https://doi.org/10.63282/3050-922X.IJERET-V2I2P108
63. Enjam, G. R., & Chandragowda, S. C. (2021). RESTful API Design for Modular Insurance Platforms. *International Journal of Emerging Research in Engineering and Technology*, *2*(3), 71-78. https://doi.org/10.63282/3050-922X.IJERET-V2I3P108
64. Pappula, K. K. (2022). Containerized Zero-Downtime Deployments in Full-Stack Systems. International Journal of AI, BigData, Computational and Management Studies, 3(4), 60-69. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I4P107
65. Jangam, S. K., Karri, N., & Pedda Muntala, P. S. R. (2022). Advanced API Security Techniques and Service Management. *International Journal of Emerging Research in Engineering and Technology*, *3*(4), 63-74. https://doi.org/10.63282/3050-922X.IJERET-V3I4P108
66. Anasuri, S. (2022). Zero-Trust Architectures for Multi-Cloud Environments. International Journal of Emerging Trends in Computer Science and Information Technology, 3(4), 64-76. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I4P107
67. Pedda Muntala, P. S. R., & Karri, N. (2022). Using Oracle Fusion Analytics Warehouse (FAW) and ML to Improve KPI Visibility and Business Outcomes. International Journal of AI, BigData, Computational and Management Studies, *3*(1), 79-88. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I1P109
68. Rahul, N. (2022). Optimizing Rating Engines through AI and Machine Learning: Revolutionizing Pricing Precision. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(3), 93-101. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I3P110
69. Enjam, G. R. (2022). Secure Data Masking Strategies for Cloud-Native Insurance Systems. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(2), 87-94. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I2P109
70. Pappula, K. K. (2023). Edge-Deployed Computer Vision for Real-Time Defect Detection. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 72-81. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I3P108
71. Jangam, S. K. (2023). Importance of Encrypting Data in Transit and at Rest Using TLS and Other Security Protocols and API Security Best Practices. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 82-91. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I3P109
72. Anasuri, S., & Pappula, K. K. (2023). Green HPC: Carbon-Aware Scheduling in Cloud Data Centers. *International Journal of Emerging Research in Engineering and Technology*, *4*(2), 106-114. https://doi.org/10.63282/3050-922X.IJERET-V4I2P111

73. Reddy Pedda Muntala , P. S. (2023). Process Automation in Oracle Fusion Cloud Using AI Agents. International Journal of Emerging Research in Engineering and Technology, 4(4), 112-119. https://doi.org/10.63282/3050-922X.IJERET-V4I4P111
74. Enjam, G. R. (2023). Optimizing PostgreSQL for High-Volume Insurance Transactions & Secure Backup and Restore Strategies for Databases. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(1), 104-111. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P112
75. Pappula, K. K., & Rusum, G. P. (2021). Designing Developer-Centric Internal APIs for Rapid Full-Stack Development. *International Journal of AI, BigData, Computational and Management Studies*, *2*(4), 80-88. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I4P108
76. Pedda Muntala, P. S. R. (2021). Integrating AI with Oracle Fusion ERP for Autonomous Financial Close. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 76-86. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I2P109
77. Jangam, S. K. (2022). Role of AI and ML in Enhancing Self-Healing Capabilities, Including Predictive Analysis and Automated Recovery. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), 47-56. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P106
78. Anasuri, S., Rusum, G. P., & Pappula, kiran K. (2022). Blockchain-Based Identity Management in Decentralized Applications. International Journal of AI, BigData, Computational and Management Studies, *3*(3), 70-81. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I3P109
79. Pedda Muntala, P. S. R. (2022). Enhancing Financial Close with ML: Oracle Fusion Cloud Financials Case Study. *International Journal of AI, BigData, Computational and Management Studies*, *3*(3), 62-69. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I3P108
80. Jangam, S. K., & Karri, N. (2023). Robust Error Handling, Logging, and Monitoring Mechanisms to Effectively Detect and Troubleshoot Integration Issues in MuleSoft and Salesforce Integrations. *International Journal of Emerging Research in Engineering and Technology*, *4*(4), 80-89. https://doi.org/10.63282/3050-922X.IJERET-V4I4P108
81. Anasuri, S. (2023). Synthetic Identity Detection Using Graph Neural Networks. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(4), 87-96. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P110
82. Reddy Pedda Muntala, P. S., & Karri, N. (2023). Voice-Enabled ERP: Integrating Oracle Digital Assistant with Fusion ERP for Hands-Free Operations. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(2), 111-120. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I2P111
83. Enjam, G. R., Tekale, K. M., & Chandragowda, S. C. (2023). Zero-Downtime CI/CD Production Deployments for Insurance SaaS Using Blue/Green Deployments. *International Journal of Emerging Research in Engineering and Technology*, *4*(3), 98-106. https://doi.org/10.63282/3050-922X.IJERET-V4I3P111
84. Pedda Muntala, P. S. R., & Jangam, S. K. (2021). End-to-End Hyperautomation with Oracle ERP and Oracle Integration Cloud. *International Journal of Emerging Research in Engineering and Technology*, *2*(4), 59-67. https://doi.org/10.63282/3050-922X.IJERET-V2I4P107
85. Jangam, S. K., & Karri, N. (2022). Potential of AI and ML to Enhance Error Detection, Prediction, and Automated Remediation in Batch Processing. *International Journal of AI, BigData, Computational and Management Studies*, *3*(4), 70-81. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I4P108