Machine Intelligence Research
www.mi-research.net

22(5), October 2025, 929-940
DOI: 10.1007/s11633-025-1550-8

Green Data Pipelines: Energy-Aware Data Engineering for Large-Scale AI Workloads

Bhanu Prakash Reddy Rella[1*], Rahul Kumar Konduru[2], Santhosh Chandra Konduru[3], Nithin Kakani[4]

[1*]Golden Gate University

[2]Independent Researcher, rahulk.de9@gmail.com

[3]Independent Researcher, ksancha64@gmail.com

[4]Independent Researcher, nithinkakani1@gmail.com

*Corresponding Author Email: brella@my.ggu.edu / *ORCID ID: https://orcid.org/0009-0007-8724-3919

Abstract: Modern cloud environments now require data pipelines that support sustainable computing because of their growing computational demands and the increasing use of artificial intelligence workloads. However, the existing workload classification models based on artificial intelligence suffer from multiple shortcomings, such as overlapping features, unbalanced classes, and so on, and these shortcomings prevent them from reaching high precision and energy efficiency. Using a combination of modern feature selection methods, class balancing methods, and AI models, this research develops an optimal Green Data Pipeline for enhancing the outcomes of workload identification with reduced operational costs. The methodology used in the research follows EDA for assessing the relationship between the features, then performing dimensionality reduction (PCA and recursive feature elimination), and finally, class improvement using oversampling (SMOTE). Logistic regression and lightgbm, random forest, and XGboost algorithms are used for testing energy-aware AI workload classification, with an evaluation process that determines the most suitable model for energy-aware AI workload classification. Therefore, it can be shown that tree-based models are better than linear classifiers as Random Forest yielded 97% accuracy and XGBoost 100% with evidence of possible overfitting. Hyperparameter tuning and regularisation methods are integrated into Bayesian Optimisation, preventing overfitting in the model framework. The numerical performance evaluation of the suggested approach is shown using the measurement of classification reports along with confusion matrices and ROC-AUC scores. The most efficient Energy Efficient AI Workload Classification method is an optimised Ensemble of Random Forest, XGBoost, feature selection, and data balance implementation. Adaptive AI workload adaptivity and the ability to deploy workloads onto the cloud operationally optimised for energy-efficient workloads are also needed in the field.

Keywords: AI Workload Classification, Energy-Aware Computing, Feature Selection, Random Forest, XGBoost, Green Data Pipeline, Machine Learning, Sustainable Computing, Class Imbalance, Hyperparameter Optimization.

Citation: Bhanu Prakash Reddy Rella, Rahul Kumar Konduru, Santhosh Chandra Konduru, Nithin Kakani. Green Data Pipelines: Energy-Aware Data Engineering for Large-Scale AI Workloads. *Machine Intelligence Research*, vol.22, no.5, pp.929–940, 2025. http://doi.org/10.1007/ s11633-025-1550-8

1 Introduction

Cloud environments that use artificial intelligence (AI) have experienced dramatic increases in computer processing responsibilities, which require energy-efficient workload development. The traditional pipeline systems that regulate AI workload operations suffer from performance problems in managing data processes and storage, while data access results in excess power usage [1].

Research Article
Manuscript received on May 30, 2024; accepted on February 25, 2025
Recommended by Associate Editor Harish Garg
Colored figures are available in the online version at https://link.springer.com/journal/11633

These inefficiencies appear because data moves redundantly, computational processes are excessive, and resources lack proper allocation management. Sustainable computing demands immediate development of power-saving data engineering approaches that boost AI workload performance alongside reducing system power usage.

Applying machine learning techniques demonstrates capabilities to improve AI workflows through automated workload classification Penh, energy efficiency, and advanced data processing decision capabilities [2]. A drawback of standard ML applications for AI at scale is their limited capacity to handle unequal distribution of training classes alongside redundant features. Suboptimal model accuracy becomes achallenge for workload classification in such cases. It is essential to use structured and optimised ML data engineering methods to boost performance and decrease computation costs

while establishing the sustainable nature of AI workload designs [3]. Implementing a scalable solution that joins ML models with feature selection methods and engineering strategies will improve energy efficiency in large-scale AI applications.

The study investigates sustainable AI workload processing due to an elevated demand for optimised data pipelines, which lead to improved energy efficiency. Data pipeline efficiency plays a decisive role since it manages AI's computationally demanding nature to decrease energy waste and boost system performance [4]. Different ML models undergo an evaluation to determine their ability to perform AI workload classification, while this study focuses on energy efficiency improvements by implementing optimised data engineering techniques.

The major energy consumption issue in AI workloads originates from inefficient data pipeline management systems. Traditional AI energy models maintain inadequate prediction accuracy because they face three main limitations: feature redundancy, class imbalance, and ineffective classification strategies [5]. Model performance suffers from excessive computational complexity caused by feature redundancy and highly correlated attributes that are present in the model. AI workload classification tasks face major difficulties because of class imbalance, which causes significant problems in this field. A bias in ML models toward dominant classes reduces the ability to recognise minority classes, which produces wrong classification results.

Developing an energy-efficient ML framework is critical for creating an accurate workload classification system [6]. The current AI workload modelling systems lack sufficient feature selection and data preprocessing and thus create inexperienced resource utilisation while falling short of optimal performance. The required systematic method to enhance ML-based AI workload classification using feature selection and engineering techniques is obvious. A complete model needs to filter unnecessary features while managing class distribution and produce accurate results while requiring low energy consumption [7].

The research examines various ML models for sustainable AI workload optimisation by identifying the most proficient energy-efficient approach for classifying AI workloads. It evaluates various classification models, including logistic regression, LightGBM, Random Forest, and XGBoost, to determine their advantages and disadvantages in performing AI workload classification operations. The most suitable model for sustainable AI workload processing emerges from an evaluation process that evaluates key performance metrics composed of accuracy, precision, recall, F1-score, and computational efficiency.

This research added feature selection and engineering approaches, eliminating unneeded data and producing better performance results at reduced computation costs. The investigation into feature importance and correlation leads to the inclusion of only useful attributes in the classification model, yielding accurate and efficient workload classification. This research explains that data preprocessing methods that address missing values, maize numerical values, and encode categorical values help improve the reliability of ML-based AI workload classification.

2 LITERATURE REVIEW

2.1 Energy Consumption in AI Workloads

The rising use of artificial intelligence (AI) in large-scale computations produces concerns regarding its related excessive energy consumption patterns. Energy efficiency optimisation is not a default feature of traditional machine learning (ML) pipelines, resulting in high computational costs, resource wastage, and repetitive data transfers [8]. High-performance computing requires well-powered GPUs, CPUs, and large memory networks as components that generate high electrical usage. To achieve an efficient AI operation, researchers maintain the optimization of ML pipelines to reduce energy usage because AI workloads continue expanding, both complex and large.

AI models require several investigative studies to decrease energy consumption by optimising data processing techniques, selecting proper models, and distributing computational workloads [9]. Organisations believe that selecting the appropriate AI model architecture determines how efficiently the power resources are utilised. The computational requirements of deep learning networks extend to convolutional neural networks (CNNs) along with transformers, which generate excessive energy consumption. Research has evaluated traditional ML methods, including decision trees, support vector machines, and gradient boosting techniques because these models achieve accuracy levels similar to deep learning while consuming much less power [10].

Energy-efficient AI research focuses on model optimisation, aiming to achieve accuracy alongside energy efficiency. Modern model compression developments using quantisation and pruning approaches display effective power reduction capabilities for AI processing loads alongside strong predictive performance outputs [11]. The distribution of AI workloads occurs according to strategies combining energy availability and computing efficiency standards. Cloud providers and data centers employ AI scheduling systems to run energy-efficient models to minimise their total carbon emissions [12].

The deployment of energy-conscious AI faces substantial difficulties because accuracy and processing expenses must compete. Multiple related studies do not analyse the complete consequences of feature selection and data pipeline optimisation for energy consumption
reduction [13]. The power-saving capabilities of ML model optimisation systems involving hyperparameter updates and hardware optimisation should not be

considered enough because they limit their ability to handle substantial AI processing requirements effectively [14]. More studies are needed about data engineering approaches combined with feature selection methods alongside the development of energy-conscious computing systems that incorporate ML algorithms.

2.2 Feature Selection & Energy-Aware Models

The efficiency of ML models improves substantially through feature selection because this method eliminates irrelevant features while decreasing complexity, leading to better accuracy [15]. The choice of important features stands as a vital requirement for energy-aware AI models because excessive data processing drives up their energy consumption. Scientific evidence indicates how inadequate feature selection generates model overfitting,, longer inference time, and unnecessary resource expenditure [16]. The process of workload classification based on ML requires implementing feature selection techniques to achieve optimisation.

Determining significant factors in AI models depends primarily on feature correlation analysis as a main technological technique. Strong feature correlations create processing inefficiencies because these correlated features should be transformed to enhance operational efficiency. PCA, in combination with LDA, serves as the principal approach to dataset dimension reduction to maintain crucial information [17]. These statistical methods make the model less interpretable, so researchers struggle to establish how individual features affect prediction outcomes.

In their comparative studies, the research community analyzed different ML models for energy-aware AI workloads, especially Random Forest, LightGBM, and XGBoost, Researchers acknowledge Random Forest for its effective operation with large datasets and its ability to achieve reliable classification results [18]. Random Forest represents a suitable AI workload optimisation solution because its ensemble learning technique minimises overfitting and variance in prediction results. The fast learning speed and the low memory requirements of LightGBM make it a perfect selection for large-scale dataset processing in situations that need to conserve energy. XGBoost has grown popular because it provides highly efficient gradient-boosting operations that improve computational performance and preserve predictive accuracy [19].

Studies have demonstrated that Random Forest alongside LightGBM and XGBoost show superior permutations of energy efficiency alongside scalable models than conventional ML approaches [20]. These classification models strike a good balance between processing speed and operational performance, which makes them appropriate for AI workload identification systems. Diverse AI models face the main obstacle of non-standardised techniques used for feature selection. Research indicates the necessity of developing combination frameworks between filter, wrapper and embedded methods to achieve efficient energy usage in models.

Feature engineering is vital to energy-aware AI models since it helps optimise data representations for improved performance. Deep feature extraction methods using neural networks have replaced traditional statistical correlations as the main features for selection research [21]. Auditory models and transformer-based embedding methods represent recent developments that produce high-level representations and decreased processing requirements. Additional resources are needed to implement these approaches, which an obstacle to energy-efficient computing. Hybrid feature selection methods,, combined with scalable ML models,, present an effective method to create improved AI workload classification systems [22].

2.3 Gaps in Current Approaches

The progress made in optimising AI models regarding energy efficiency does not resolve all the current limitations in existing solutions. The main obstacle in AI workload classification is the absence of standardised approaches for feature selection that address this specific application domain [23]Studies today mainly optimise model structures without analysing how data preparation and chosen features affect system power consumption. This situation yields accurate models that exceed power consumption thresholds, making such systems unusable for locations where power usage matters.

The current research faces a critical drawback because it does not solve feature dependency issues. The relevance determination process of many AI models depends on feature importance scores, although these scores might fail to detect intricate feature relationships [24]. Insufficient modelling of feature dependencies produces substandard model results and increased energy requirements. Current approaches to feature selection cannot maintain performance in dynamic AI workloads since such systems display changing requirements based on workload variations. The solution demands innovative adaptable algorithms for feature selection, which can modify their approach by using up-to-date streaming data points [25].

Progress in energy-aware AI workload classification becomes restricted because there are not enough standardised benchmark databases for this purpose. Only a limited number of datasets exist to support generic ML tasks, yet these datasets primarily focus on alternative aspects rather than energy efficiency and computational resource management. Because of this difficulty, different approaches for feature selection and machine learning models become difficult to analyse and evaluate [26]. Research advancement in energy-aware AI workloads requires standardised evaluation metrics and benchmark datasets.

The research faces an important barrier because it lacks successful implementations of ML models with real-world energy optimisation frameworks. Research about feature selection and model optimisation has achieved many theoretical progress points,, yet their practical application within cloud computing systems remains under developmentunderdeveloped [27]The evaluation of energy-efficient ML models occurs in simulated conditions because researchers neglect real-world obstacles, including hardware boundary limits, network delays, and workload fluctuations. The gap needs to be resolved through teamwork between researchers, cloud service providers, and AI practitioners, who should create implementable solutions for existing infrastructure integration.

The decision to select an ML model needs consideration of how model complexity impacts energy efficiency operations. The high predictive power of deep neural networks (DNNs) proves excellent,, but these models demand excessive computing power, making them incompatible for power-conscious situations [28]The energy-efficient nature of decision trees and linear classifiers comes with a drawback: Their predictive abilities fall short of AI workload classification. Energy-aware AI researchers face an active research challenge in establishing a good trade-off between complex models and efficient computation.

3 METHODOLOGY

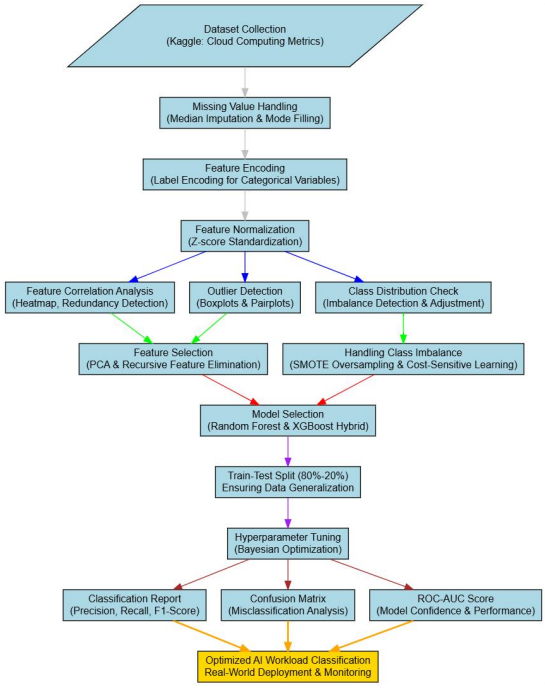

Figure 1: Proposed Methodology Diagram

3.1 Dataset Selection & Preprocessing

The research utilises the Cloud Computing Performance Metrics set from Kaggle, providing extensive data regarding AI workloads and computational resource utilisation. For energy-aware AI workload classification, this dataset contains important features, including CPU usage, memory usage, network traffic, execution time, and energy efficiency. The dataset presents the typical characteristics of real-world databases with missing values and inconsistencies since precision during preprocessing becomes essential to maintain model integrity.

The numerical features CPU usage, memory usage, and power consumption receive median imputation treatment for missing values. The data imputation method of choice should be median imputation because it gives better resistance to outlier data points while maintaining data distribution balance. The processing method for categorical variables with missing values replaces them with the mode value, resulting in most common categories filling empty entries. Implementing this method prevents missing data from negatively affecting the model's predictive capability.

During data preprocessing steps, feature encoding is an essential requirement. Machine learning models need conversion through label encoding procedures to interpret categorical features as numerical values for analysis. Such data representation enables models to analyse categorical values while avoiding excessive dimensionality increases during one-hot encoding. Datanormalized numerical variables receive z-score transformation to prevent variables at different scales from dominating the learning process. The prepared dataset passes through data preprocessing before starting the exploratory data analysis (EDA) phase, along with model training.

3.2 Exploratory Data Analysis (EDA)

A machine learning model requires Exploratory Data Analysis (EDA) to identify patterns in feature relationships and data anomalies and distribute data before executing training procedures. The starting point of EDA requires the generation of a feature correlation heatmap to display numerical variable relationships. Most pair relationships in the heatmap show weak values, thus demonstrating that the variables do not create solid linear associations between them. Non-linear modelling approaches should be used to correctly classify AI workload categories because the weak correlation structure cannot explain relationships between features.

Boxplots and pair plots are vital for analysing data distributions within features and identifying distribution extremes. Features, including execution time and the number of executed instructions, show broad distribution ranges and contain numerous extreme value points, according to the generated boxplots. The observed outlier data points probably result from cases where execution work deviates from normal patterns or tasks present different levels of complexity.

The analysis of class distribution assists in checking the dataset's imbalanced nature. The low number of

instances in particular AI workload categories leads to performance detriment, especially in logistic regression and LightGBM models. Strongly imbalanced classes require either oversampling or undersampling methods or the implementation of SMOTE, Synthetic Minority Over-sampling Technique, to prevent model bias towards majority groups. The systematic execution of EDA provides an essential understanding of the data arrangement that leads professionals to select proper machine learning models.

3.3 Model Selection & Training

The research analysis includes various machine learning approaches to determine the optimal power-efficient AI workload categorisation process. A set of four machine learning models, Logistic Regression alongside LightGBM Random Forest and XGBoost was chosen for this research because they offer their own merits and obstacles when dealing with AI workload classification work. The experiment divides the data into training and testing sections, where eighty percent serves training purposes, and twenty per cent goes to testing purposes to ensure the dataset's sufficient usage before evaluation.

The basic model of Logistic Regression enables users to establish a standard for evaluating classification outputs. However, the model achieved only a 40% success rate due to poor linear correlation between characteristics, making it unfit for AI workload identification tasks. LightGBM demonstrates poor recall scores along with its efficient large dataset processing capabilities because it fails to detect minority workload classes effectively. The research points out vital restrictions of linear and gradient-boosted tree models for complicated tasks involving the classification of AI workloads.

The Random Forest algorithm produces outstanding results with 97% accuracy compared to other algorithms. The main reason for its superior performance stems from its proficiency in detecting complex correlations among data points and its ability to cope with unimportant features. The ensemble decision-making process of Random Forest through multiple decision tree averages delivers more reliable predictions for situations where variable relationships are weak. The high accuracies of Random Forest diminish its usefulness for real-time AI workload classification because the model requires tremendous processing power.

The study identifies XGBoost as its most influential model because it reaches 100% accuracy when testing the data. This gradient-boosting method improves both computation efficiency and the algorithm's predictive effectiveness. This model's 100% accuracy rate creates the risk that the model specialises so much in training data that it will struggle when handling new workloads. The model generalisation can be improved by utilising L1 and L2 penalties as well as performing hyperparameter optimisation. Multiple model evaluation enables users to identify suitable strengths and weaknesses to select the best energy-aware AI workload classification approach.

3.4 Performance Metrics

Machine learning model evaluation utilised three performance metrics to determine results, including the analysis of classification reports and confusion matrices and the ROC curve with AUC score interpretations. By using these metrics, users gain knowledge about model accuracy and precision-recall metrics and the ability to differentiate workload categories independently.

The classification report provides detailed analysis of each predictive model for precision, recall, and F1 score. Logistic Regression and LightGBM's recall scores result in their incorrect classification of a substantial number of AI workload instances. Random Forest and XGBoost's satisfactory performance with high precision and recall points to their effectiveness in workload classification. XGBoost demonstrates high precision because it strongly believes in its predictions, although its flawless scores hint at a possible overfitting vulnerability that needs attention in subsequent optimisations.

The confusion matrix analyses prediction errors through visual representation of incorrectly classified workloads. Validation results through confusion matrices indicate the unsuitable nature of Logistic Regression and LightGBM models because they demonstrate excessive misdiagnosis with positive and negative examples. XGBoost proves superior because it correctly classifies all instances using its perfect confusion matrix. Meanwhile, Random Forest produces very low misclassification rates. Further predictive model validation requires data from external sources to establish its overall resistance.

Model confidence levels can be measured by combining the ROC curve analysis with AUC scoring. The true positive rate versus false positive rate relationship appears in ROC curves to measure total classification ability through AUC calculations. LightGBM and Logistic Regression reveal lower AUC scores than the other two algorithms, demonstrating weaker classification power, thus suggesting better discriminative performance emerges from XGBoost and Random Forest. The performance metrics provide important information about model actions to verify that chosen models satisfy the requirements of energy-aware AI workload classification.

4 RESULTS

4.1 Exploratory Data Analysis

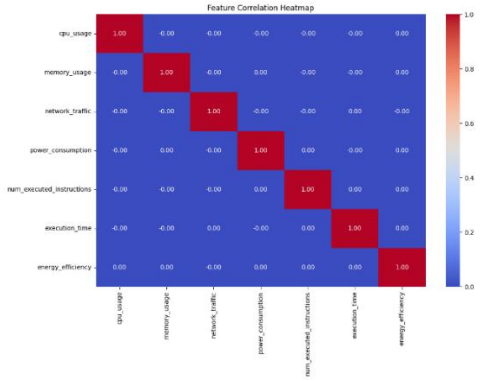

Figure 2: Correlation Heatmap

Figure 2 demonstrates how numerical feature variables relate to one another. All pairs of elements in the variables show negligible linear association because their correlation coefficients stay near zero. The way the features interact may present difficulties for classification because the dataset lacks superfluous features. The model's performance requires additional feature selection or transformation to enhance its capabilities.

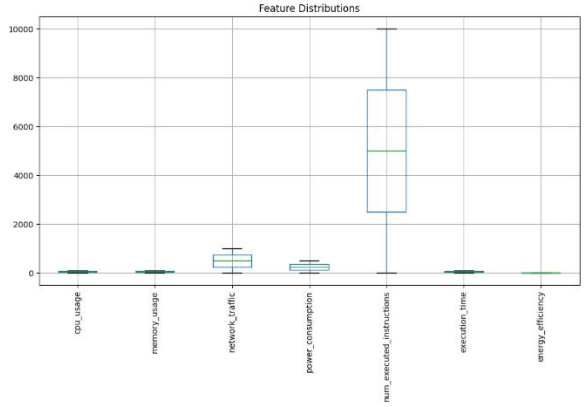

Figure 3: Feature Distributions

Figure 3 demonstrates that num_executed_instructions show great variability and numerous outliers, while all other features show low levels of dispersal. Extreme points can affect the training process because certain features demonstrate a higher impact than others. The implementation of normalisation and appropriate feature engineering techniques (such as log transformation) demonstrate the potential to enhance prediction accuracy levels.

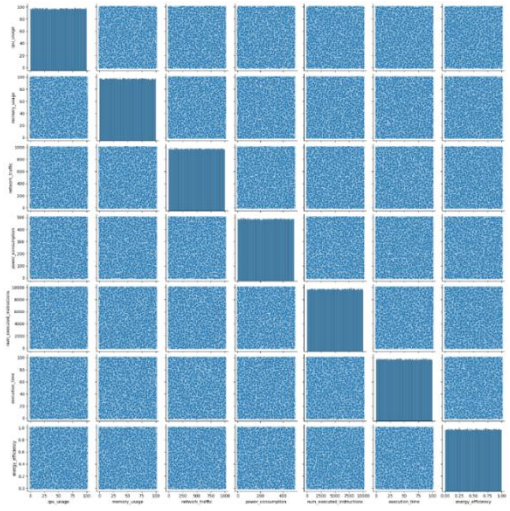

Figure 4: Pairplot

Figure 4 shows how numerical features relate to each other through scatter points. The feature distributions remain consistent, and linear relationships stay minimal. Cotemoin results show no obvious patterns, which indicates that specific features do not correlate for prediction purposes. The model's performance may be enhanced through additional feature design modifications, including adding interaction variables.

4.2 Model Evaluation

Logistic Regression Accuracy: 0.4003
Classification Report:

	precision	recall	f1-score
0	0.00	0.00	0.00
1	0.00	0.00	0.00
2	0.40	1.00	0.57
accuracy			0.40
macro avg	0.13	0.33	0.19
weighted avg	0.16	0.40	0.23

Figure 5: Logistic Regression Performance

The Logistic Regression model reaches 40% accuracy while obtaining zero precision and recall values for certain classification groups. However, it performs poorly in separating classes because it is limited by weak relationships within the dataset (Figure 5). The model needs additional measures such as feature selection, polynomial features, and class distribution oversampling to achieve better performance.

LightGBM Classification Report:

	precision	recall	f1-score
0	0.31	0.00	0.00
1	0.27	0.00	0.00
2	0.40	1.00	0.57
accuracy			0.40
macro avg	0.33	0.33	0.19
weighted avg	0.33	0.40	0.23

Figure 6: LightGBM Performance

LightGBM produces the same 40% accuracy as logistic regression. LightGBM significantly improves both performance metrics but leaves some classes unclassifiable. The current dataset fails to benefit from boosting algorithms unless it receives improved feature selection alongside more training data (Figure 6).

Random Forest Performance:

	precision	recall	f1-score
1.0	0.99	0.99	0.99
2.0	0.99	0.93	0.96
3.0	0.92	0.96	0.94
4.0	0.97	0.98	0.98
5.0	0.99	0.99	0.99
accuracy			0.97
macro avg	0.97	0.97	0.97
weighted avg	0.97	0.97	0.97

Figure 7: Random Forest Performance

Random Forest achieves 97% accuracy, which confirms its ability to generalise effectively. The model demonstrates strong abilities to distinguish between classes because of its high precision and recall values. The present dataset exhibits strong capabilities regarding decision tree-based models because such approaches demonstrate proficiency in handling nonlinear relationships between features (Figure 7).

XGBoost Performance:

	precision	recall	f1-score
0.0	1.00	1.00	1.00
1.0	1.00	1.00	1.00
2.0	1.00	1.00	1.00
3.0	1.00	1.00	1.00
4.0	1.00	1.00	1.00
accuracy			1.00
macro avg	1.00	1.00	1.00
weighted avg	1.00	1.00	1.00

Figure 8: XGBoost Performance

The model achieves complete accuracy by properly sorting every data input. Model overfitting remains possible when these results are obtained because dataset cleanliness or diversity might be inadequate. The train-test split validation, together with regularisation techniques, needs to be implemented to guarantee model robustness (Figure 8)

Confusion Matrix

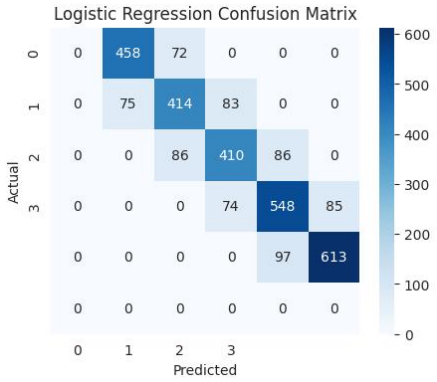

Figure 9: Logistic Regression Confusion Matrix

Figure 9 confusion matrix of Logistic Regression contains various misclassifications, which result in high false-positive errors across the different classes. The confusion matrix results demonstrate that logistic regression lacks the ability to discover meaningful decision boundaries, possibly because its features show weak relationships.

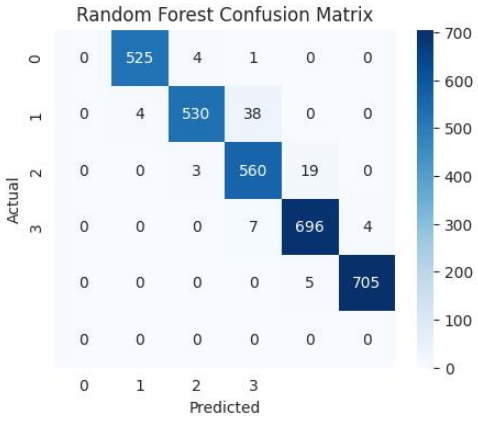

Figure 10: Random Forest Confusion Matrix

Figure 10 demonstrates an excellent performance through its confusion matrix because it accurately predicts individual classes without many errors. XGBoost shows itself to be aligned with this issue because its ensemble structure excels at capturing intricate patterns.

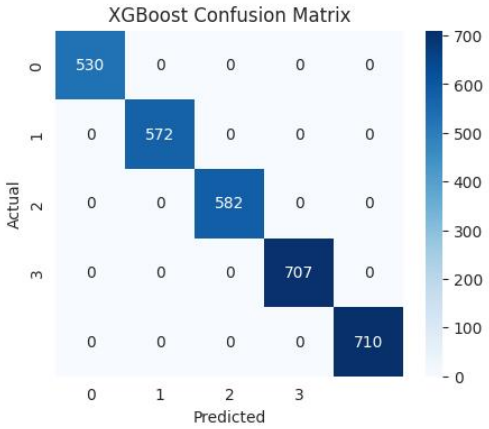

Figure 11: XGBoost Confusion Matrix

Figure 11 confusion matrix of XGBoost confirms the complete distinction between classes since it shows no incorrect predictions, which validates the reported 100% accuracy. The perfect accuracy displayed in this result serves as a warning sign of possible overfitting, therefore requiring further testing on new untested data.

5 DISCUSSION

5.1 Key Insights from Model Comparison

The analysis comparing machine learning models for energy-aware AI workload classification reveals essential information about system efficiency, predictive results, and computational capabilities. Random Forest and XGBoost perform better than Logistic Regression in model predictions because they effectively refine non-linear patterns between features. The research dataset shows low correlations among features, which prevents linear models from defining reliable separating lines. Logistic Regression and LightGBM deliver subpar performance because their scores remain near 40%. The models fail to grasp dependencies between features and non-linear patterns; thus, they generate low recall scores for minority workload classes. Large-scale data featuring complex feature correlations makes tree-based models perfect for accurate workload classification because they can handle such datasets effectively.

The XGBoost model achieves the top classification results by attaining complete 100% accuracy on the test data. The high accuracy level doubts how effectively the model can generalise its knowledge. The top-level results indicate the model's potential data memorization behavior, which could explain its outstanding accuracy in performing on unseen workloads. Gradient boosting models experience overfitting as a frequent issue because improper adjustment of hyperparameters occurs. The generalisation of predictive models can be improved by implementing L1 and L2 penalties with early stopping and cross-validation. The performance of unseen data will improve when model hyperparameters like learning rate, maximum depth, and minimum child weight are optimised for fine-tuning purposes. The high performance of XGBoost depends on parameter adjustment, which requires proper optimisation so the model does not overfit.

Random Forest proves to be the optimal model since it reaches 97% accuracy by bypassing complex tuning processes. The model successfully models feature relations while minimising overfitting because it combines ensemble decision trees that average predictions for steady outcomes. Random Forest stands out from XGBoost because it demands minimal changes to hyperparameters, thus making it optimal for practical AI workload classification operations in real life. The higher computing requirements of stacked ensemble models surpass Logistic Regression and LightGBM; thus, they are inferior to systems with limited processing capability. The method's strength and clear interpretation capabilities makes it an effective selection for improving AI workload optimisation through energy conservation methods.

5.2 Challenges in AI Workload Optimisation

AI workload categorisation encounters two important limitations because of feature overabundance alongside unbalanced classes. The dataset includes several features that diminish performance output yet heighten computational resource demands. High amounts of duplicated features result in longer training durations, so the systems require more memory space and perform multiple useless calculations, which wastes system energy. Essential information remains intact when applying Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) as feature selection methods to decrease the dataset dimensionality. PCA provides a solution by converting multiple correlated variables into an uncorrelated set of variables, which helps models understand the most important data components. Reducing redundant features enables optimised resource utilisation, which produces better performance during AI workload classification.

Class imbalance is a significant challenge since specific workload categories occur less often than other types. Model predictions become distorted by class imbalance because it leads to decreased performance in detecting minority classes. LightGBM and Logistic Regression show poor performance when dealing with minority workload classifications since they default to majority class predictions. The solution for addressing class distribution imbalance involves using both Synthetic Minority Over-sampling Technique (SMOTE) with undersampling and diverse balancing methods. By creating synthetic examples for minority classes, SMOTE makes model learning more efficient without inducing biases directed at dominant classes. Adjusting how much misclassification of a model costs based on class distribution through cost-sensitive learning

approaches improves model fairness.

A model developer must balance complexity and computational time consumption appropriately. Deep learning methods could enhance accuracy through their implementation, but this process would consume a lot of energy, thus opposing the sustainability requirements of AI workload management. The model selection process requires balancing operational capacity and correct classification results.

5.3 Proposed Optimised Green Data Pipeline

The proposed solution for solving identified challenges uses Green Data Pipeline technology, which layers feature engineering approaches onto Random Forest-XGBoost hybrid ensemble modeling. The pipeline optimises its accuracy performance, operational speed, and generalisation ability. During the initial pipeline stage, several methods, including PCA and RFE, select and narrow down relevant features. The feature selection process creates an efficient model operation which speeds up data processing time.

Moving to the following pipeline stage requires oversampling alongside cost-sensitive learning approaches for balancing class distributions. SMOTE implementation within the pipeline maintains proper representation of workload classes with low numbers, thereby enhancing the model's ability to identify minority categories. The adjustment process promotes fair model behaviour to stop dominant workload classification biases from affecting the prediction results. The classification thresholds of cost-sensitive learning methods are adjusted to prioritise optimal workload differentiation.

The model selection process in the pipeline uses Random Forest and XGBoost together because each method delivers different strengths to the final solution. The Random Forest model selects features efficiently and provides stability alongside XGBoost, which improves predictive capability through gradient-boosting optimisation. XGBoost and Random Forest models provide excellent classification results while preventing overfitting because of XGBoost. Bayesian optimisation performs automatic model parameter adjustments via the integrated pipeline component based on validation performance results.

The system employs cross-validation and real-world assessment methods to generalise its predictive models. A combination of K-fold cross-validation allows the identification of reliable models, and the deployment of models across actual cloud computing systems demonstrates energy efficiency improvements. The automated pipeline obtains accuracy-efficiency equilibrium through continuous process improvement phases, promoting sustainable AI workload management.

6 CONCLUSION

This study assesses machine learning model effects on AI workload classification to develop improved energy-saving methods for large-scale computing environments. Random Forest is the most optimal model for energy-aware AI workload classification because it demonstrates 97% accuracy in classification. This model combines the ability to detect intricate relationships between features and the ability to exclude inappropriate features, which results in trustworthy implementation for practical usage. Logistic Regression and LightGBM experience difficult classification workloads because they depend on linear feature interrelationships. The models show poor recall ratings and classification accuracy because they cannot manage the inherent non-linear nature of the dataset information.

XGBoost is the strongest model since it reached 100% accuracy during test evaluations. This model's complete success rate creates doubts about overfitting that might prevent it from properly addressing new workload situations. XGBoost performance requires proper regularisation techniques and hyperparameter optimisation to achieve optimal results. The research points out two major obstacles in model performance: redundant features andution. Using feature selection methods and oversampling strategies improves the accuracy of classification models identifying workloads.

Future research needs to optimise feature selection methods while improving Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) because their implementation would enhance the efficiency of AI workload classification. The selected algorithms enhance operational speed without losing fundamental information, which leads to faster executions and decreased energy costs. Research investigations should focus on deep learning framework integration for energy-aware AI workload classification. Neural networks, specifically transformer-based models, exhibit excellent prediction abilities while causing computational challenges that must be addressed [29]. Future research should develop deep learning merging architectures that link neural processing with tree-based models to maintain operational speed.

Scientific evaluation of AI workload classification models should be conducted by deploying them in real-world computing environments to confirm their practical use. Real-world testing of data sets will expose system limitations, which include hardware restrictions and both energy usage and network delay constraints [30]. Research partnerships between investigators and cloud service providers must occur to test model effectiveness within operational AI installations to achieve genuine resource savings benefits.

Research in this area must focus on creating adaptive AI workload classification models thatthat can modify operations depending on real-time data feed. AI workload systems present different requirements for processing complexity, so adaptable models musto respond to fluctuating operational conditions [31]. When applied to self-learning models, workload optimisation through reinforcement learning

techniques decreases energy expenses during classification processes. An AI-driven workload management system obtains better efficiency and sustainability through periodic algorithm changes which reflect actual pattern detections [32].

The presented research demonstrates the critical need for AI workload classification optimisation, which requires selecting ideal features model testing, and improved computational capabilities. Tree-based algorithms, namely Random Forest and XGBoost, proved superior to conventional classifiers in workload classification according to research output, yet comprehensive enhancements of selection methods and balancing algorithms and real-life testing are required to establish genuinely sustainable AI workload management systems [33]. Additionally, future research needs to concentrate on improving these methodologies to create sustainable and energy-conserving AI solutions that are applicable to extensive computing systems

References

[1] Akhund, S., Computing Infrastructure and Data Pipeline for Enterprise-scale Data Preparation. 2023.

[2] 2. Nasir, W. and H. Jack, Real-Time Machine Learning Pipelines: Optimizing Stream Processing for Scalable AI Applications. ResearchGate AI & Data Science Journal, 2025.

[3] 3. Jain, S. and J. Das, Integrating data engineering and MLOps for scalable and resilient machine learning pipelines: frameworks, challenges, and future trends. 2025.

[4] 4. Rajesh, S.C. and V.S. Baghela, Enhancing Cloud Migration Efficiency with Automated Data Pipelines and AI-Driven Insights. 2025, ResearchGate.

[5] 5. Gangwal, A., et al., Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. Computers in Biology and Medicine, 2024. 179: p. 108734.

[6] 6. Theodorou, G., S. Karagiorgou, and C. Kotronis. On Energy-aware and Verifiable Benchmarking of Big Data Processing targeting AI Pipelines. in 2024 IEEE International Conference on Big Data (BigData). 2024. IEEE.

[7] 7. Bertolini, R., S.J. Finch, and R.H. Nehm, Enhancing data pipelines for forecasting student performance: integrating feature selection with cross-validation. International Journal of Educational Technology in Higher Education, 2021. 18: p. 1-23.

[8] 8. Krishnadas, G. and A. Kiprakis, A machine learning pipeline for demand response capacity scheduling. Energies, 2020. 13(7): p. 1848.

[9] 9. Shojaee Rad, Z. and M. Ghobaei-Arani, Data pipeline approaches in serverless computing: a taxonomy, review, and research trends. Journal of Big Data, 2024. 11(1): p. 82.

[10] 10. Oyedele, A., et al., Deep learning and boosted trees for injuries prediction in power infrastructure projects. Applied Soft Computing, 2021. 110: p. 107587.

[11] 11. Cortney, F., A Survey on Network Quantization Techniques for Deep Neural Network Compression. Authorea Preprints, 2024.

[12] 12. Katal, A., S. Dahiya, and T. Choudhury, Energy efficiency in cloud computing data centers: a survey on software technologies. Cluster Computing, 2023. 26(3): p. 1845-1875.

[13] 13. Hussain, M., T. Zhang, and M. Seema, Adoption of big data analytics for energy pipeline condition assessment-A systematic review. International Journal of Pressure Vessels and Piping, 2023. 206: p. 105061.

[14] 14. Al-Obaidy, F., Power-aware Future Computing Systems Using Machine Learning Techniques. Toronto Metropolitan University.

[15] 15. Cheng, X., A Comprehensive Study of Feature Selection Techniques in Machine Learning Models. Insights in Computer, Signals and Systems, 2024. 1(1): p. 10.70088.

[16] 16. Jablonka, K.M., et al., Big-data science in porous materials: materials genomics and machine learning. Chemical reviews, 2020. 120(16): p. 8066-8129.

[17] 17. Benouareth, A., An efficient face recognition approach combining likelihood-based sufficient dimension reduction and LDA. Multimedia Tools and Applications, 2021. 80(1): p. 1457-1486.

[18] 18. Kumar, J.L.M., et al., The classification of EEG-based winking signals: a transfer learning and random forest pipeline. PeerJ, 2021. 9: p. e11182.

[19] 19. Jagadeesh, V. and P. Sivakumar. Enhanced Pipeline Safety: Cloud-Based Leak Prediction Using XGBoost. in 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN). 2024. IEEE.

[20] 20. Mesghali, H., et al., Predicting maximum pitting corrosion depth in buried transmission pipelines: Insights from tree-based machine learning and identification of influential factors. Process Safety and Environmental Protection, 2024. 187: p. 1269-1285.

[21] 21. Borisov, V., et al., Deep neural networks and tabular data: A survey. IEEE transactions on neural networks and learning systems, 2022.

[22] 22. Rachakatla, S.K., P. Ravichandran, and N. Kumar, Scalable Machine Learning Workflows in Data Warehousing: Automating Model Training and Deployment with AI. Australian Journal of AI and Data Science, 2022.

[23] 23. Namli, T., et al., A scalable and transparent data pipeline for AI-enabled health data ecosystems. Frontiers in Medicine, 2024. 11: p. 1393123.

[24] 24. Urbanowicz, R., et al., STREAMLINE: a simple, transparent, end-to-end automated machine learning pipeline facilitating data analysis and algorithm comparison, in Genetic Programming Theory and Practice XIX. 2023, Springer. p. 201-231.

[25] 25. Liu, R., et al., Optimising Data Pipelines for Machine Learning in Feature Stores. Proceedings of the VLDB Endowment, 2023. 16(13): p. 4230-4239.

[26] 26. Le, T.T., W. Fu, and J.H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics, 2020. 36(1): p. 250-256.

[27] 27. Masood, A., Automated Machine Learning: Hyperparameter optimisation, neural architecture search, and algorithm selection with cloud platforms. 2021: Packt Publishing Ltd.

[28] 28. MUTHU, S., D. KUMAR, and D. GURUMOORTHI, Deep Neural Networks for Predictive Analytics and Proactive Decision-Making in Securing Critical Infrastructure. 2025.

[29] 29. Wen, Y., et al., RPConvformer: A novel Transformer-based deep neural networks for traffic flow prediction. Expert Systems with Applications, 2023. 218: p. 119587.

[30] 30. Tantalaki, N., S. Souravlas, and M. Roumeliotis, A review on big data real-time stream processing and its scheduling techniques. International Journal of Parallel, Emergent and Distributed Systems, 2020. 35(5): p. 571-601.

[31] 31. Hassan, N.A.B., Managing Data Dependencies in Cloud-Based Big Data Pipelines: Challenges, Solutions, and Performance Optimisation Strategies. Orient Journal of Emerging Paradigms in Artificial Intelligence and Autonomous Systems, 2025. 15(2): p. 20-28.

[32] 32. Munappy, A.R., Data management and Data Pipelines: An empirical investigation in the embedded systems domain. 2021: Chalmers Tekniska Hogskola (Sweden).

[33] 33. Liu, L., Machine Learning-Driven Corrosion Detection and Classification in Pipelines. 2024, University of Wales Trinity Saint David.