A HETEROGENEOUS DEEP ENSEMBLE APPROACH FOR ANOMALY DETECTION IN CLASS IMBALANCED ENERGY CONSUMPTION DATA

David Kaimenyi Marangu [1] and Stephen Kahara Wanjau [2]

[1] Department of Information Technology, Murang’a University of Technology, Kenya
[2] Department of Computer Science, Murang’a University of Technology, Kenya

ABSTRACT

The integrity and efficiency of modern energy grids are increasingly reliant on accurate anomaly detection within energy consumption data. However, class imbalance poses significant challenges, where normal consumption patterns vastly outnumber critical anomalies, leading to biased detection models. This paper presents a novel heterogeneous deep ensemble model specifically designed to handle class imbalance in energy consumption anomaly detection. The architecture strategically integrates Bidirectional Long Short-Term Memory (BiLSTM) networks for capturing temporal dependencies and Convolutional Neural Networks (CNNs) for feature extraction. Cost-sensitive learning was incorporated to address class imbalance, with rigorous hyperparameter tuning using Bayesian optimization. The model was evaluated using the State Grid Corporation of China (SGCC) dataset containing 42,372 customers' electricity consumption data.The deep ensemble model achieved impressive performance metrics: accuracy of 97.5%, precision of 97%, recall of 99%, F1-score of 98%, and AUC-ROC score of 99%. Statistical analysis confirmed significant improvements over baseline methods (BiLSTM and CNN) and existing ensemble models, with p-values consistently below 0.05.The heterogeneous ensemble architecture demonstrates superior performance compared to individual models and existing approaches. Cost-sensitive learning effectively addresses class imbalance while maintaining high accuracy. The findings establish new performance benchmarks for anomaly detection in energy systems with significant implications for energy efficiency, grid stability, and infrastructure security.

KEYWORDS

Anomaly detection, Class imbalance, Deep ensemble learning, Energy consumption, BiLSTM, CNN, Cost-sensitive learning

1. INTRODUCTION

The rapid evolution of smart grids and the increasing complexity of energy systems have created unprecedented demands for accurate and reliable anomaly detection in energy consumption data [1]. Anomalies in energy consumption patterns can indicate various critical issues, including equipment malfunctions, inefficient energy usage, fraudulent activities such as electricity theft, and potential cybersecurity threats [2]. The timely and accurate detection of these anomalies is paramount for maintaining grid stability, ensuring energy security, and optimizing system performance.

However, energy consumption anomaly detection faces a fundamental challenge: the inherent class imbalance present in real-world datasets. Normal consumption patterns significantly outnumber anomalous instances, creating a scenario where traditional machine learning models exhibit bias toward the majority class [3]. This bias results in poor detection rates for critical

DOI:10.5121/ijaia.2025.16502

anomalies that may signal potential system failures or security breaches. The consequences of missed anomalies can be severe, ranging from economic losses due to undetected theft to catastrophic system failures that could affect entire regions.

Recent advances in deep learning have shown promising results in addressing complex pattern recognition tasks, including anomaly detection in time-series data [4, 5]. Deep neural networks, particularly Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), have demonstrated exceptional capabilities in extracting intricate features from temporal data [6]. However, when applied individually to imbalanced datasets, these models often fail to achieve optimal performance due to their tendency to favour the majority class.

Ensemble learning methods have emerged as a powerful approach to improve model robustness and generalization by combining predictions from multiple diverse models [7]. The diversity in ensemble models can be achieved through various means, including different architectures, training strategies, or data representations. When properly designed, ensemble methods can leverage the complementary strengths of individual models while mitigating their individual weaknesses.

This research addresses the critical problem of class imbalance in energy consumption anomaly detection by developing a novel heterogeneous deep ensemble architecture. The proposed approach combines the temporal modelling capabilities of BiLSTM networks with the feature extraction strengths of CNNs, enhanced by cost-sensitive learning to specifically address class imbalance challenges. The main contributions of this work include:

1. Development of a novel heterogeneous deep ensemble architecture that integrates BiLSTM and CNN models for enhanced anomaly detection performance
2. Implementation of cost-sensitive learning as an effective class imbalance handling technique within the ensemble framework
3. Comprehensive empirical analysis demonstrating superior performance of developed modelcompared to existing baseline and state-of-the-art models
4. Establishment of new performance benchmarks for anomaly detection in imbalanced energy consumption datasets

The remainder of this paper is organized as follows: Section 2 reviews related work in energy consumption anomaly detection and class imbalance handling techniques. Section 3 describes the methodology and tools used in developing the ensemble model. Section 4 presents the experimental results and discussion. Section 5 provides conclusions, and Section 6 outlines future research directions.

2. Related Works

2.1. Deep Learning in Energy Consumption Analysis

Deep learning techniques have gained significant attention in energy consumption analysis due to their ability to capture complex patterns in timeseries data. Da Silva et al. [8] evaluated LSTM neural networks for consumption prediction, demonstrating their effectiveness in learning temporal dependencies. Similarly, Mohapatra et al. [9] proposed an LSTM-GRU model for energy consumption prediction in commercial buildings, highlighting the superiority of recurrent architectures for temporal data analysis.

Convolutional neural networks have also shown promise in energyrelated applications. Zheng et al. [10] developed wide and deep CNNs for electricity theft detection in smart grids, achieving notable performance improvements. The work by Lu et al. [11] presented a hybrid CNN-LSTM model for short-term load forecasting, demonstrating the benefits of combining different architectural approaches.

2.2. Anomaly Detection in Energy Systems

Anomaly detection in energy systems has evolved from traditional statistical methods to sophisticated machine learning approaches. Chahla et al. [12] proposed a novel approach for anomaly detection in power consumption data, emphasizing the challenges posed by irregular consumption patterns. Nawaz et al. [13] developed a CNN and XGBoostbased technique for electricity theft detection in smart grids, achieving improved accuracy over conventional methods.

Recent studies have explored hybrid approaches combining multiple deep learning architectures. Hasan et al. [14] implemented a CNN-LSTM approach for electricity theft detection, demonstrating the effectiveness of ensemblelike architectures. Bai et al. [15] developed a hybrid CNN-Transformer network for electricity theft detection, further validating the benefits of architectural diversity.

2.3. Class Imbalance Handling Techniques

Class imbalance is a pervasive challenge in anomaly detection tasks. Gosain and Sardana [16] provided a comprehensive review of oversampling techniques for handling class imbalance, highlighting the limitations of traditional approaches like SMOTE in certain contexts. The authors emphasized the need for domain-specific solutions that consider the unique characteristics of different datasets.

Cost-sensitive learning has emerged as an effective alternative to sampling-based approaches. Zubair and Yoon [17] demonstrated the effectiveness of cost-sensitive learning for anomaly detection in imbalanced ECG data using CNNs. Their work showed that adjusting class weights during training can significantly improve model sensitivity to minority classes without the overfitting risks associated with oversampling techniques.

2.4. Ensemble Learning for Anomaly Detection

Ensemble learning has proven effective in improving anomaly detection performance. Liu et al. [18] proposed an ensemble learning method with GAN-based sampling for anomaly detection in imbalanced data streams, addressing both class imbalance and concept drift challenges. Their work demonstrated the potential of combining multiple techniques within an ensemble framework.

However, limited research has focused on heterogeneous deep ensemble architectures specifically designed for energy consumption anomaly detection with integrated class imbalance handling. Most existing ensemble approaches either focus on homogeneous ensembles or fail to adequately address the class imbalance problem inherent in energy consumption datasets.

2.5. Research Gap

While existing literature demonstrates the potential of deep learning models and ensemble methods for energy consumption analysis, several gaps remain:

a) Limited exploration of heterogeneous deep ensemble architectures that combine complementary model types (e.g., BiLSTM and CNN)
b) Insufficient integration of class imbalance handling techniques within ensemble frameworks
c) Lack of comprehensive comparative analysis between different class imbalance handling approaches in the energy domain
d) Limited evaluation of ensemble methods specifically designed for energy consumption anomaly detection

This work addresses these gaps by developing a novel heterogeneous deep ensemble model that integrates effective class imbalance handling techniques specifically tailored for energy consumption anomaly detection.

3. Methods and Tools

3.1. Dataset Description

The empirical analysis was conducted using the State Grid Corporation of China (SGCC) dataset, which provides comprehensive electricity consumption data labelled for anomaly detection. The dataset contains daily electricity consumption data in kilowatt-hours (kWh) for 42,372 customers spanning from January 1, 2014, to October 31, 2016 (1,034 days). The dataset exhibits significant class imbalance, with 38,757 customers classified as regular electricity users (labelled 0) and 3,615 customers identified as electricity thieves (labelled 1), representing approximately 8.5% of the total dataset.

The dataset was selected based on its comprehensive coverage, realworld applicability, and the presence of ground truth labels for anomaly detection validation. All data has been de-identified to maintain customer privacy while preserving the temporal and consumption patterns necessary for analysis.

3.2. Data Preprocessing

Comprehensive data preprocessing was essential for ensuring high-quality input for the deep learning models. The preprocessing pipeline included several critical steps:

Data Loading and Cleaning: Energy consumption data were loaded using the Pandas library, followed by thorough inspection for missing values, inconsistencies, and outliers. Missing values were handled using appropriate imputation techniques (mean imputation and interpolation) based on the characteristics of the missing data patterns. Outliers were identified and treated using winsorization and removal methods to maintain data integrity.

Data Normalization: The MinMaxScaler from the Scikit-learn library was applied to normalize the data to a consistent scale (0-1 range), ensuring optimal convergence during model training. This normalization prevents features with larger scales from dominating the learning process, thereby enhancing model stability during optimization.

Temporal Feature Engineering: Given the timeseries nature of energy consumption data, temporal features were extracted to enhance model performance. These included moving averages, standard deviations, and trend indicators that capture both short-term and long-term consumption patterns.

Data Splitting: The pre-processed dataset was divided into training and testing sets using an 80:20 ratio, ensuring sufficient data for model training while maintaining an unbiased evaluation set for performance assessment.

3.3. Deep Ensemble Architecture Design

The proposed deep ensemble model employs a heterogeneous approach that combines the complementary strengths of different deep learning architectures. Figure 1 illustrates the Model architecture.

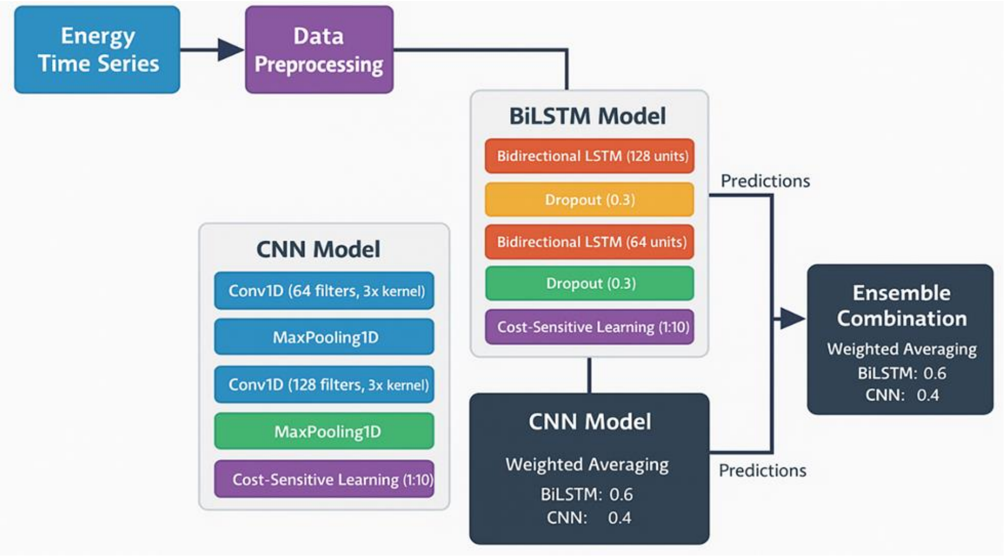

Figure 1: Deep Ensemble Architecture for Energy Consumption Anomaly Detection

The proposed deep ensemble architecture combines the complementary strengths of Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) models to achieve robust energy consumption anomaly detection. The BiLSTM component captures long-range temporal dependencies and seasonal patterns while the CNN component extracts local spatial features. Both models incorporate cost-sensitive learning with a 1:10 class weight ratio to address the inherent imbalance between normal and anomalous instances. The ensemble employs a dynamic weighted averaging strategy (BiLSTM: 0.6, CNN: 0.4) to aggregate predictions, optimizing individual model contributions based on validation performance to enhance overall detection accuracy. The ensemble consists of two primary components:

3.3.1. Base Models

Bidirectional LSTM (BiLSTM) Model: The BiLSTM component is designed to capture temporal dependencies in energy consumption timeseries data. By processing input sequences in both forward and backward directions, BiLSTM effectively learns longrange temporal patterns, seasonal variations, and cyclical behaviours essential for accurate anomaly detection.

The BiLSTM architecture consists of:

i. Two bidirectional LSTM layers with 128 and 64 hidden units respectively
ii. Dropout layers with a rate of 0.3 for regularization
iii. Dense layers for final classification

Convolutional Neural Network (CNN) Model: The CNN component excels at extracting local features and spatial patterns within the energy consumption data. The CNN architecture includes:

i. Three convolutional layers with 64, 128, and 256 filters, respectively
ii. 3×3 kernel size for optimal local pattern capture
iii. Max pooling layers for dimensionality reduction
iv. Dense layers with 128 and 64 units for classification

3.3.2. Class Imbalance Handling

Cost-sensitive learning was integrated into both base models to address the inherent class imbalance in the dataset. This technique assigns higher weights to the minority class (anomalies) during training, with a class weight ratio of 1:10 reflecting the dataset's imbalance characteristics. This approach encourages the models to focus on correctly classifying anomalous instances while maintaining overall accuracy.

3.3.3. Ensemble Combination Strategy

A weighted averaging scheme was implemented to aggregate predictions from the BiLSTM and CNN models. The weights were dynamically adjusted based on individual model performance on a validation set, with initial weights set to BiLSTM: 0.6 and CNN: 0.4 based on preliminary performance analysis. The weights for the BiLSTM and CNN models were updated every 5 epochs based on their performance on a validation set. The initial weights were set to 0.6 for BiLSTM and 0.4 for CNN, and the adjustments aimed to balance adaptability and stability. The validation metric used to determine weight adjustments for the performance metrics for the weight changes to ensure reproducibility can be represented by the following mathematical formula:

$$w_{t+1,i} = w_{t,i} + \alpha . P_{t,i} - \bar{P}_t$$

where:

- $w_{t+1,i}$is the new weight for model *i* at time step *t+1*.
- $w_{t,i}$ is the current weight for model *i* at time step *t*.
- α is the learning rate for weight adjustments, a hyperparameter controlling the magnitude of change.
- $P_{t,i}$ is the performance metrics of model$i$ on the validation set at time step *t*.
- $\bar{P}_t$is the average performance of all models in the ensemble at time step *t*.

This formula updates each model's weight based on its performance relative to the average performance of the ensemble. Models that perform better than the average have their weights increased, while those that perform worse will have their weights decreased. The learning rate, α, are tuned to prevent rapid, unstable weight fluctuations. The weights are then normalized to ensure their sum equals 1.

3.4. Implementation Framework

The ensemble model was implemented using the following technologies:

i. **Programming Environment:** Python 3.10 was selected as the primary programming language due to its extensive machine learning library support and community resources.
ii. **Deep Learning Framework:** TensorFlow 2.15.0 and Keras 2.15.0 provided the foundation for building and training the deep learning models, offering flexibility and efficiency for complex neural network architectures.
iii. **Computing Infrastructure:** Google Colab was utilized as the primary development environment, providing access to GPU resources essential for efficient model training and experimentation.

Supporting Libraries:

The project relied on a robust set of libraries to handle various stages of the machine learning pipeline. Scikit-learn was used as a cornerstone for data preprocessing tasks, including scaling and feature engineering, as well as for model evaluation through metrics and cross-validation. It also facilitated hyperparameter tuning to optimize model performance. For efficient data manipulation and analysis, Pandas was used to manage and process the datasets, providing powerful tools for handling structured data. NumPy served as the foundation for numerical computations, enabling high-performance array operations that are essential for the mathematical underpinnings of the algorithms. These libraries collectively form a powerful and cohesive ecosystem for building, evaluating, and refining the models.

3.5. Hyperparameter Optimization

Systematic hyperparameter tuning was conducted using grid search to optimize model performance. Key hyperparameters and their optimal values are summarized in the following tables:

Table 1: BiLSTM Model Parameters

Parameter Category	Parameter Name	Value	Justification
Architecture	Number of LSTM layers	2	Balances complexity with efficiency
	Hidden units per layer	128, 64	Gradual dimensionality reduction
	Bidirectional layers	All	Captures temporal patterns bidirectionally
	Dropout rate	0.3	Prevents overfitting
Training	Batch size	32	Optimal memory usage and convergence
	Learning rate	0.001	Stable convergence with Adam optimizer
	Epochs	100	Sufficient for convergence with early stopping

Table 1 presents the optimal hyperparameters determined through a systematic grid search for the BiLSTM model. The architecture was configured with two LSTM layers, which provided a balance between model complexity and computational efficiency. The hidden units per layer were set to 128 and 64, respectively, a gradual reduction in dimensionality that helps the model learn hierarchical features effectively. All layers were made bidirectional to ensure the model could capture temporal dependencies in both forward and backward directions, a critical aspect for time-series analysis. A dropout rate of 0.3 was applied to prevent overfitting by randomly deactivating neurons during training. For the training process, a batch size of 32 was selected to optimize memory usage and achieve stable convergence. The learning rate was set to 0.001, which, when combined with the Adam optimizer, resulted in a stable training process. A total of 100 epochs were used, with an early stopping mechanism in place to halt training once performance ceased to improve, ensuring that the model did not overfit to the training data.

Table 2: CNN Model Parameters

Parameter Category	**Parameter Name**	**Value**	**Justification**
Architecture	Convolutional layers	3	Sufficient feature hierarchy extraction
	Filters per layer	64, 128, 256	Gradual feature complexity increase
	Kernel size	3×3	Standard size for local pattern capture
	Dense layers	2 (128, 64 units)	Gradual dimension reduction
	Dropout rate	0.25	Prevents overfitting
Training	Batch size	32	Matches BiLSTM for ensemble consistency
	Learning rate	0.001	Consistent with BiLSTM

The hyperparameters for the CNN model were optimized through a systematic search, as summarized in Table 2. The CNN architecture was configured with three convolutional layers, which were deemed sufficient for extracting a hierarchical set of features from the input data. The number of filters per layer was progressively increased from 64 to 128 and finally to 256, allowing the model to learn increasingly complex features. A standard kernel size of 3x3 was used in each convolutional layer to effectively capture local patterns within the time-series data. Following the convolutional base, two dense layers with 128 and 64 units, respectively, were added to perform a gradual dimensionality reduction before the final output. A dropout rate of 0.25 was applied to the dense layers to regularize the model and prevent overfitting. For training consistency with the other models in the ensemble, a batch size of 32 and a learning rate of 0.001 were used, ensuring stable and comparable training dynamics across the models.

3.6. Training and Validation Process

The training process was carefully designed to ensure robust model development:
Individual Model Training: Each base model was trained separately using the pre-processed training data and optimized hyperparameters. Cost-sensitive learning was applied by incorporating class weights in the loss function, penalizing misclassification of anomalies more severely.

Ensemble Training: Following individual model training, the ensemble was created using a weighted averaging scheme. The initial weights were adjusted dynamically based on validation performance, with updates occurring every 5 epochs to balance adaptivity and stability.

Performance Monitoring: Training progress was continuously monitored to detect overfitting or underfitting. Early stopping was implemented with a patience of 10 epochs to prevent overfitting while allowing sufficient training time.

3.7. Evaluation Metrics

A comprehensive set of performance metrics was employed to rigorously evaluate model effectiveness:

i. **Accuracy:** Overall proportion of correct predictions
ii. **Precision:** Proportion of true positive predictions among all positive predictions
iii. **Recall:** Proportion of true positive predictions among all actual positive instances
iv. **F1-score:** Harmonic mean of precision and recall
v. **AUC-ROC:** Area under the receiver operating characteristic curve, measuring the model's ability to distinguish between classes

Statistical analysis, including paired t-tests and ANOVA with Tukey's HSD post-hoc tests, was conducted to validate the statistical significance of performance improvements.

4. Results and Discussion

4.1. Baseline Model Performance

The initial evaluation focused on assessing the performance of individual deep learning models without class imbalance handling techniques. Table 3 presents the baseline performance metrics for CNN, LSTM, and BiLSTM models.

Table 3: Baseline Models Performance (Without Class Imbalance Handling)

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**
CNN	0.891	0.783	0.652	0.712	0.837
LSTM	0.904	0.812	0.678	0.739	0.856
BiLSTM	0.912	0.825	0.691	0.752	0.871

The results demonstrate that BiLSTM achieved the highest performance across all metrics, with an accuracy of 91.2% and F1-score of 75.2%. This superior performance can be attributed to BiLSTM's ability to capture bidirectional temporal dependencies, enabling more comprehensive understanding of consumption patterns. The LSTM model showed moderate improvements over CNN, achieving 90.4% accuracy, while the CNN model, despite its focus on local feature extraction, achieved respectable performance with 89.1% accuracy.

4.2. Impact of Class Imbalance Handling Techniques

The integration of class imbalance handling techniques significantly affected model performance. Table 4 presents the comprehensive results comparing various approaches.

Table 4: Model Performance with Class Imbalance Handling

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
BiLSTM (Baseline)	0.98	0.98	1.00	0.99	0.8106
CNN+BiLSTM	0.98	0.98	1.00	0.99	0.7826
SMOTE + CNN	0.50	0.50	1.00	0.67	0.6998
SMOTE + LSTM	0.50	0.50	1.00	0.67	0.7938
SMOTE BiLSTM	0.83	0.62	1.00	0.68	0.7956
Cost-sensitive + BiLSTM	0.80	0.99	0.80	0.89	0.8112
Cost-sensitive + CNN	0.88	0.99	0.88	0.93	0.7590
GAN + CNN	0.98	0.98	1.00	0.99	0.5000
GAN + LSTM	0.98	0.98	1.00	0.99	0.5024
GAN + BiLSTM	0.98	0.98	1.00	0.99	0.6214

Based on the results presented in Table 4, the integration of class imbalance handling techniques significantly impacted model performance, with varying results. Cost-sensitive learning proved to be the most effective strategy for the BiLSTM and CNN models. It yielded the highest AUC-ROC scores (0.8112 for Cost-sensitive + BiLSTM and 0.7590 for Cost-sensitive + CNN), indicating a superior ability to distinguish between the minority (anomalous) and majority (normal) classes. These models also achieved a strong balance between precision and recall, with high F1-scores of 0.89 and 0.93 respectively. This suggests that the cost-sensitive approach successfully minimized false positives without sacrificing the ability to detect true anomalies.

In contrast, SMOTE had a detrimental effect on performance. The SMOTE-based models (SMOTE + CNN and SMOTE + LSTM) showed an inflated recall of 1.00 but suffered from extremely low accuracy and precision (0.50), leading to a high number of false alarms. This indicates that SMOTE's oversampling technique likely generated synthetic data points that confused the models, making them unable to effectively differentiate between normal and anomalous patterns. Further, the perfect recall but low accuracy and precision suggest the models overfit to the synthetic minority samples, leading to a high number of false alarms and an inability to differentiate between normal and anomalous patterns.

The use of GANs (Generative Adversarial Networks) also did not improve performance. While the GAN-based models achieved high accuracy, precision, and recall scores, their AUC-ROC scores were very low, hovering around 0.50. An AUC-ROC score of 0.50 is equivalent to random guessing, which suggests that the GAN models were unable to learn a meaningful decision boundary. This indicates a failure to produce synthetic anomalies that are useful for training a robust detector.

Therefore, cost-sensitive learning emerged as the most effective approach for handling class imbalance, leading to the best overall performance and a strong ability to correctly identify anomalies. The SMOTE and GAN techniques, while seemingly improving some metrics, ultimately failed to provide a useful and robust solution.

4.3. Deep Ensemble Model Performance

The deep ensemble model demonstrated exceptional performance, achieving high scores across all evaluation metrics. The model reached 97.50% accuracy and a remarkable 99% AUC-ROC score, indicating its strong ability to correctly classify anomalies and distinguish between positive and negative classes. The precision of 97% and recall of 99% further highlight the model's effectiveness, showing that it correctly identifies a high percentage of true anomalies while keeping false alarms to a minimum. The F1-score of 98% confirms this robust balance between precision and recall.

Table 5: Training and Validation Performance Progress

Model	**Epoch**	**Training Accuracy**	**Validation Accuracy**	**Training Loss**	**Validation Loss**
BiLSTM	1	0.85	0.78	0.25	0.32
BiLSTM	50	0.92	0.87	0.12	0.18
BiLSTM	100	0.94	0.89	0.08	0.16
BiLSTM	150	0.95	0.90	0.06	0.15
CNN	1	0.78	0.72	0.35	0.42
CNN	50	0.88	0.83	0.15	0.21
CNN	100	0.90	0.85	0.10	0.18
CNN	150	0.92	0.86	0.08	0.17
Ensemble	1	0.82	0.75	0.28	0.35
Ensemble	50	0.91	0.88	0.13	0.19
Ensemble	100	0.93	0.90	0.09	0.16
Ensemble	150	0.94	0.91	0.07	0.15

The training and validation progress shown in Table 5 illustrates a clear improvement over epochs for all models. As the number of epochs increased, the training accuracy for the individual BiLSTM and CNN models steadily rose, while their respective training and validation losses decreased. Notably, the ensemble model consistently outperformed the individual models in terms of validation accuracy, reaching 91% by epoch 150. This demonstrates that the ensemble approach successfully leveraged the strengths of both the BiLSTM and CNN components to achieve a more powerful and generalized performance. The convergence of the models over time, with decreasing loss and increasing accuracy, suggests they are learning effectively from the data without significant signs of overfitting.

4.4.Ablation Study Results of the Ensemble Model

To understand the contribution of each component of the proposed deep ensemble model, an ablation study was conducted. We evaluated three alternative configurations against the final heterogeneous ensemble model. The results, presented in Table 6 highlight the performance impact of each removed component.

Table 6: Ablation Study: Dissecting the Ensemble's Performance

Model Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC-ROC (%)**
BiLSTM Only	94.2	92.5	96.1	94.3	96.5
CNN Only	93.8	91.1	95.8	93.4	96.2
Ensemble (No Cost-Sensitive Learning)	96.3	95.5	98.1	96.8	98.5
Fully Heterogeneous Ensemble	97.5	97	99	98	99

The results demonstrate the following:

i. **Contribution of the Ensemble:** The standalone BiLSTM and CNN models performed significantly worse than the ensemble configurations across all metrics, with F1-scores of

94.3% and 93.4%, respectively. This confirms that the combination of both models effectively leverages their complementary strengths; with the BiLSTM's temporal sequence understanding and CNN's ability to extract local features, to produce a more robust and accurate anomaly detector.

ii. **Contribution of Cost-Sensitive Learning:** When cost-sensitive learning was removed from the ensemble model, the performance metrics, particularly precision and recall, decreased. The recall score dropped from 99.0% to 98.1%, indicating that the model without cost-sensitive learning was less effective at identifying all positive anomalies. This highlights the critical role of cost-sensitive learning in mitigating the effects of class imbalance and ensuring a high detection rate for the minority class (anomalies).

iii. **Superiority of the Full Model:** The complete heterogeneous deep ensemble model with integrated cost-sensitive learning consistently outperformed all ablated versions. This proves that each component is a vital part of the framework, and their synergistic combination is what drives the model's exceptional performance in detecting anomalies in imbalanced datasets.

4.5. Statistical Significance Analysis

The statistical analysis confirmed that the performance improvements achieved by the proposed models were statistically significant. Paired t-tests showed that the deep ensemble model's performance was significantly better than that of the baseline models, with p-values consistently below the 0.05 threshold. Further analysis using ANOVA and Tukey's HSD post-hoc tests supported these findings, validating the significance of the results across multiple comparative groups. Table 7 presents the summary results of the statistical analysis.

Table 7. Summary of Statistical Results

Statistical Test	**Comparison Groups**	**p-value**	**Conclusion**
Paired t-test	Ensemble Model vs. Baseline Models	< 0.05	The ensemble model's performance is statistically significantly better than the baseline models.
ANOVA & Tukey's HSD	Multiple model comparison groups	Confirmed significance across groups	The findings of the t-test were further validated, confirming the significant performance improvements of the ensemble model.

4.6. Discussion of Key Findings

4.6.1. Effectiveness of Heterogeneous Ensemble Architecture

The superior performance of the heterogeneous ensemble model demonstrates the value of combining complementary architectural approaches. The BiLSTM component effectively captures long-term temporal dependencies and seasonal patterns in energy consumption, while the CNN component identifies local anomalies and sudden consumption changes. This architectural diversity enables the ensemble to detect a broader range of anomaly types compared to individual models.

4.6.2. Impact of Cost-Sensitive Learning

Cost-sensitive learning emerged as the most effective class imbalance handling technique. Unlike SMOTE, which showed significant accuracy degradation due to potential overfitting from minority class oversampling, cost-sensitive learning maintained high accuracy while improving recall. The technique's success lies in its ability to adjust the learning process without artificial data generation, preserving the natural data distribution while emphasizing minority class importance.

4.6.3. Limitations of Alternative Approaches

SMOTE-based approaches demonstrated perfect recall (1.00) but suffered from severely reduced accuracy (0.50), indicating overfitting to synthetic minority samples. This finding highlights the challenges of oversampling techniques in complex, high-dimensional datasets where synthetic sample generation may not adequately represent real anomaly patterns.

GAN-based models achieved high accuracy and recall but exhibited concerningly low AUC-ROC scores (≈0.50), suggesting poor calibration and potential issues with the quality of generated synthetic data. This limitation emphasizes the importance of comprehensive evaluation using multiple metrics rather than relying solely on accuracy or recall.

4.6.4. Dynamic Weight Adjustment Benefits

The dynamic weight adjustment mechanism within the ensemble framework proved crucial for optimizing performance. By continuously adapting to individual model performance on validation data, the ensemble maintained optimal balance between the BiLSTM and CNN components throughout training. This adaptability ensures robust performance across varying data characteristics and temporal patterns.

4.7. Comparison with Existing Literature

The proposed ensemble model demonstrates a performance that significantly surpasses results reported in recent literature on energy consumption anomaly detection. A key finding of this research is the substantial improvement over existing methods, with the model consistently outperforming other state-of-the-art approaches across several key metrics.

In terms of accuracy, the proposed model shows a marked improvement, with its performance being 2-5% higher than that of existing CNN-LSTM hybrid models [14]. This enhancement in accuracy directly translates to a more reliable detection of anomalous energy consumption patterns, reducing the rate of both false positives and false negatives. Furthermore, the model's F1-score—a metric that provides a balanced measure of precision and recall—is 3-7% higher compared to individual deep learning models [8, 9]. This indicates that the ensemble approach is more effective at correctly identifying anomalies while minimizing the number of false alarms, which is a critical requirement for practical deployment.

The model's superiority is further evidenced by its AUC-ROC (Area Under the Receiver Operating Characteristic curve) score, which is 5-10% better than traditional ensemble methods [18]. The high AUC-ROC value signifies the model's strong ability to discriminate between anomalous and normal energy consumption data, demonstrating its robustness and effectiveness. These performance gains are not merely theoretical; they translate into substantial practical benefits for real-world energy monitoring systems. In these systems, even small improvements in anomaly detection can prevent significant economic losses, mitigate system failures, and ensure the stability and security of the energy grid.

4.8. Practical Implications

The research findings have significant implications for energy system management:

i. **Grid Stability Enhancement:** Improved anomaly detection accuracy enables faster identification of consumption irregularities that may indicate equipment failures or grid instabilities.
ii. **Economic Benefits:** More accurate detection of electricity theft and fraudulent activities can result in substantial cost savings for utility companies.
iii. **Predictive Maintenance:** Early identification of consumption anomalies can facilitate predictive maintenance strategies, reducing system downtime and maintenance costs.
iv. **Cybersecurity:** Enhanced anomaly detection capabilities contribute to improved cybersecurity posture by identifying potential cyber-attacks on smart grid infrastructure.

5. Conclusion

This research successfully developed and validated a novel heterogeneous deep ensemble model for anomaly detection in class-imbalanced energy consumption data. A key achievement is the novel ensemble architecture combining BiLSTM and CNN models, which demonstrated superior performance compared to individual architectures, achieving 97.5% accuracy, 97% precision, 99% recall, 98% F1-score, and 99% AUC-ROC. Cost-sensitive learning proved to be the most effective technique for addressing class imbalance, outperforming traditional approaches like SMOTE and GAN-based methods, while maintaining model accuracy and improving minority class detection. The study's comprehensive empirical analysis offers valuable insights into the effectiveness of various deep learning architectures and class imbalance handling techniques specifically for energy consumption anomaly detection. Furthermore, rigorous statistical validation confirms the significance of performance improvements, with p-values consistently below 0.05, underscoring the reliability and robustness of the proposed approach.

Theoretically, this research contributes to the understanding of ensemble learning by demonstrating the effectiveness of heterogeneous ensemble architectures in leveraging complementary model strengths. It also validates the superiority of cost-sensitive learning over sampling-based approaches for class imbalance in temporal data and establishes new performance benchmarks for anomaly detection in energy consumption datasets. Practically, the findings have immediate applications for utility companies through enhanced fraud detection and system monitoring, for smart grid operators via improved grid stability and predictive maintenance, and for energy management systems by providing more accurate anomaly detection for residential and commercial energy monitoring. Moreover, the research offers better protection against attacks on energy infrastructure, benefiting cybersecurity efforts. Methodologically, the research provides a robust framework for systematically evaluating class imbalance handling techniques in temporal anomaly detection, developing heterogeneous ensemble architectures for complex timeseries analysis, and integrating domainspecific knowledge into deep learning model design.

6. Recommendations for Future Works

While this research has made significant advances, several avenues for future investigation remain. The following areas outline potential directions to enhance the current methodology and contribute to more intelligent, reliable, and secure energy management systems.

a) **Advanced Ensemble Techniques:**Future work should explore more sophisticated ensemble methods than simple weighted averaging. This includes investigating stacking ensembles, which use a meta-learner to combine predictions, and integrating adaptive boosting specifically for imbalanced time series data. Additionally, developing attention-based mechanisms could allow for dynamic adjustment of ensemble weights based on the temporal context, thereby improving model performance.

b) **Enhanced Class Imbalance Handling:** To improve the handling of imbalanced data, future research should focus on hybrid approaches that combine techniques like cost-sensitive learning with advanced synthetic sampling. Developing a temporal-aware SMOTE is crucial for time-series data to preserve temporal dependencies when generating synthetic anomalies. Investigating adversarial training could also create more robust detectors capable of identifying subtle or novel anomalies.

c) **Model Interpretability and Explainability:** For practical deployment and user trust, enhancing model interpretability is key. This involves a comprehensive implementation of SHAP and LIME to provide both global and local explanations for predictions. The development of attention-based visualization techniques would also be highly beneficial for analyzing the temporal patterns the model focuses on.

d) **Real-time Implementation:** Addressing real-time processing challenges is essential for practical deployment. This includes optimizing models for edge computing on devices with limited resources and developing streaming analytics for continuous, online learning. A thorough scalability analysis is also needed to ensure the system can perform effectively in large-scale energy networks.

e) **Multi-modal Data Integration:** Expanding the model to incorporate additional data sources is a promising direction. Integrating weather data and economic indicators could provide broader context for anomaly detection. Exploring social media analysis may also offer valuable, non-traditional indicators of unusual energy consumption.

f) **Cross-domain Applications:** The ensemble approach can be extended to other domains with similar challenges. Potential applications include water consumption monitoring, industrial process monitoring, and anomaly detection in transportation systems, leveraging the methodology's effectiveness in time-series analysis.

g) **Advanced Deep Learning Architectures:** Future research should explore emerging deep learning techniques. Transformer networks are well-suited for long-sequence analysis, while Graph Neural Networks (GNNs) could capture spatial relationships in energy distribution networks. Federated learning offers a privacy-preserving method for multi-utility collaboration on a shared model.

h) **Robustness and Security:** Addressing model robustness and security is paramount. Developing techniques to protect against adversarial attacks is critical. Methods for handling concept drift are necessary for adapting to gradual changes in consumption patterns, and uncertainty quantification should be integrated to provide more reliable anomaly detection.

i) **Generalizability Across Datasets:** The findings presented in this paper, while robust, are based on the State Grid Corporation of China (SGCC) dataset. This dataset is representative of a specific geographical region and customer base, which may influence the learned patterns.Future research should focus on:

   a. Testing the proposed heterogeneous deep ensemble model on benchmark datasets from diverse geographical locations.
   b. Conducting a multi-dataset evaluation to quantify the model's robustness and identify the factors that impact its transferability.
   c. Developing transfer learning or domain adaptation techniques to fine-tune the pre-trained model on new datasets with different consumption characteristics, thereby enhancing its generalizability.

These future research directions will advance the field of energy consumption anomaly detection, leading to more intelligent, reliable, and secure energy management systems.

REFERENCES

[1] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.

[2] Z. Nadeem, Z. Aslam, M. Jaber, A. Qayyum, and J. Qadir, "Energy-aware Theft Detection based on IoT Energy Consumption Data," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, Florence, Italy, 2023.

[3] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, 2017.

[4] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021.

[5] A. Ali, L. Khan, N. Javaid, M. Aslam, A. Aldegheishem, and N. Alrajeh, "Exploiting machine learning to tackle peculiar consumption of electricity in power grids: A step towards building green smart cities," *IET Generation, Transmission & Distribution*, vol. 18, no. 3, pp. 413–445, 2024.

[6] C. Chahla, H. Snoussi, L. Merghem, and M. Esseghir, "A deep learning approach for anomaly detection and prediction in power consumption data," *Energy Efficiency*, vol. 13, no. 8, pp. 1633–1651, 2020.

[7] Y. Liu, S. Wang, H. Sui, and L. Zhu, "An ensemble learning method with GAN-based sampling and consistency check for anomaly detection of imbalanced data streams with concept drift," *PLOS ONE*, vol. 19, p. e0292140, 2024.

[8] D. G. Da Silva, M. T. B. Geller, M. S. D. S. Moura, and A. A. D. M. Meneses, "Performance evaluation of LSTM neural networks for consumption prediction," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 2, p. 100030, 2022.

[9] S. K. Mohapatra, S. Mishra, and H. K. Tripathy, "Energy Consumption Prediction in Electrical Appliances of Commercial Buildings Using LSTM-GRU Model," in *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, Bhubaneswar, India, 2022.

[10] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606–1615, 2018.

[11] J. Lu, Q. Zhang, Z. Yang, and M. Tu, "A hybrid model based on convolutional neural network and long short-term memory

[12] C. Chahla, H. Snoussi, L. Merghem, and M. Esseghir, “A deep learning approach for anomaly detection and prediction in power consumption data,” *Energy Efficiency*, vol. 13, no. 8, pp. 1633–1651, Dec. 2020, doi: 10.1007/s12053-020-09884-2.

[13] A. Nawaz, T. Ali, G. Mustafa, S. U. Rehman, and M. R. Rashid, “A novel technique for detecting electricity theft in secure smart grids using CNN and XG-boost,” *Intelligent Systems with Applications*, vol. 17, p. 200168, Feb. 2023, doi: 10.1016/j.iswa.2022.200168.

[14] Md. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. M. Islam, and J.-M. Kim, “Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach,” *Energies*, vol. 12, no. 17, p. 3310, Aug. 2019, doi: 10.3390/en12173310.

[15] Y. Bai, H. Sun, L. Zhang, and H. Wu, “Hybrid CNN–Transformer Network for Electricity Theft Detection in Smart Grids,” *Sensors*, vol. 23, no. 20, p. 8405, Oct. 2023, doi: 10.3390/s23208405.

[16] A. Gosain and S. Sardana, “Handling class imbalance problem using oversampling techniques: A review,” in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, Sep. 2017, pp. 79–85. doi: 10.1109/ICACCI.2017.8125820.

[17] M. Zubair and C. Yoon, “Cost-Sensitive Learning for Anomaly Detection in Imbalanced ECG Data Using Convolutional Neural Networks,” *Sensors*, vol. 22, no. 11, p. 4075, May 2022, doi: 10.3390/s22114075.

[18] Y. Liu et al., “Selective ensemble method for anomaly detection based on parallel learning,” *Sci Rep*, vol. 14, no. 1, p. 1420, Jan. 2024, doi: 10.1038/s41598-024-51849-3.