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

David Kaimenyi Marangu

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Information Technology of Murang'a University of Technology

July, 2025

DECLARATION

Student Declaration

I hereby declare that this thesis is my original work and to the best of my knowledge has not been presented for a degree awarded in this or any other university. I understand what plagiarism is, and I am aware of the university policy in this regard. I understand that any false claim in respect of this work shall result in disciplinary action in accordance with the University Anti-Plagiarism Policy

_____ _____

David Kaimenyi Marangu Date
SC500/5100/2017

Supervisors' Approval

The undersigned certify that they have read and hereby recommended for acceptance of Murang'a University of Technology thesis entitled **"A Deep Ensemble Model for Anomaly Detection in Class Imbalanced Energy Consumption Data".**

_____ _____

Dr. Stephen Thiiru Njenga, Date
Department of Computer Science,
Murang'a University of Technology.

_____ _____

Dr Rachael Njeri Ndung'u, Date
Department of Information Technology,
Murang'a University of Technology.

DEDICATION

This work is dedicated to my lovely wife, Lucy, and our children, Mervyn and Brayden.

You have been an inspiration to me.

ACKNOWLEDGEMENT

Acknowledging divine support, I express gratitude to God for His sustaining grace throughout this doctoral journey.

My profound gratitude goes to Murang'a University of Technology Vice Chancellor Prof. Dickson M. Nyariki and the University management for their generous financial sponsorship for my PhD program in Information Technology at Murang'a University of Technology. This sponsorship has been invaluable in offsetting tuition fees for the course. Your support is greatly appreciated.

I am particularly indebted to my supervisors, Dr Stephen Thiiru Njenga and Dr Rachael Njeri Ndung'u, for their tireless efforts in guiding and exposing me to research writing. Thank you for your patience, understanding, and exceptional balance of guidance and freedom, which is unmatched. Thank you for your motivation and perseverance to read the drafts, not forgetting the papers.

I would also like to thank the School of Computing and Information Technology (SCIT) fraternity at Murang'a University of Technology for the excellent research insights they supplied during my studies. Thank you for your helpful comments during the concept, proposal and seminar presentations.

My sincere gratitude also goes to my family, my wife Lucy, and my children Mervyn and Brayden. Thank you for your prayers, moral support and encouragement. You have been a pillar in these studies. Special gratitude also goes to my loving parents and supportive family who stood with me throughout the PhD journey.

To every person whom God has put along my path, I appreciate you greatly.

ABSTRACT

The integrity and efficiency of modern energy grids are increasingly reliant on the ability to detect anomalies within energy consumption data accurately. However, a significant challenge persists: the inherent class imbalance. Normal consumption patterns vastly outnumber rare, yet critical, anomalies, leading to a detrimental bias in conventional anomaly detection models. These models often prioritize the majority class, yielding poor detection rates for the vital anomalies that signal potential system failures or inefficiencies. The primary objective of this research was to develop and evaluate a deep ensemble model for anomaly detection in imbalanced energy consumption data. The specific objectives were formulated in this study as follows: Investigate existing models used to handle class imbalance in anomaly detection for energy consumption, with the aim of identifying effective techniques that can inform model selection for empirical analysis; conduct an empirical analysis of class imbalance handling models in the domain of anomaly detection for energy consumption; develop a deep ensemble learning model that integrate cost sensitive learning, a class imbalance handling technique, for anomaly detection for energy consumption data and evaluate the performance of the newly developed deep ensemble model in addressing class imbalance challenge in the domain of anomaly detection for energy consumption. A heterogeneous deep ensemble architecture was meticulously crafted to overcome the existing approaches limitations. This architecture strategically integrates the strengths of Bidirectional Long Short-Term Memory (BiLSTM) networks, capable of capturing temporal dependencies, and Convolutional Neural Networks (CNNs), proficient in feature extraction. This ensemble was further fortified by the integration of Cost-sensitive Learning, an advanced class imbalance handling technique that ensured a more balanced representation of anomaly patterns. Rigorous hyperparameter tuning, employing Bayesian optimization, was undertaken to optimize the performance of the constituent models and the overall ensemble. This study adopted a pragmatist philosophical stance, emphasizing practical outcomes and the real-world effectiveness of deep ensemble models. A comprehensive suite of performance metrics, encompassing accuracy, recall, precision, AUC-ROC, and F1-score, was employed to rigorously evaluate the model's efficacy. Statistical analysis, including paired t-tests and ANOVA with Tukey's HSD post-hoc tests, was conducted to validate the statistical significance of the results. To further enhance the understanding and trust in the model, SHAP and LIME-based explainability techniques were used to increase model transparency and interpretability. Specifically, the deep ensemble model attained an impressive accuracy of 97.5%, a precision of 97%, a recall of 99%, an F1-score of 98%, and an AUC-ROC score of 99%. The empirical results indicate that the deep ensemble model achieved consistently higher performance metrics compared to baseline methods (BiLSTM and CNN) as well as relevant cutting-edge ensemble models for anomaly detection in energy consumption data advanced by Huang et al. and Gao et al., suggesting enhanced efficacy in anomaly detection in imbalanced datasets. The statistical analysis further confirmed the significance of these improvements, with p-values consistently below 0.05 when comparing the proposed ensemble model to baseline models and other cutting-edge models. This research contributes to a novel heterogeneous deep ensemble learning architecture, demonstrates the efficacy of cost-sensitive learning in class imbalance handling and establishes new performance benchmarks for anomaly detection in energy systems.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
LIST OF TABLES ..... **x**
LIST OF FIGURES ..... **xi**
ACRONYMS AND ABBREVIATIONS ..... **xii**
CHAPTER ONE: INTRODUCTION **1**
1.1 Background of the Study 1
1.2 Problem Statement 6
1.3 Objectives 8
1.3.1 General Objective 8
1.3.2 Specific Objectives 8
1.4 Research Questions 8
1.5 Research Hypothesis 9
1.6 Significance of the Study 9
1.7 Scope of the Study 10
1.8 Limitations of the Study 10
1.9 Contributions of the Thesis 11
1.10 Organization of the Thesis 12
CHAPTER TWO: LITERATURE REVIEW **14**
2.1 Introduction 14
2.2 Theoretical Review 14
2.2.1 Computational Theory 15
2.2.2 Extreme Value Theory 18
2.2.3 Systems Theory 19
2.2.4 Statistical Learning Theory 21
2.2.5 Traditional Machine Learning Algorithms 23
2.2.6 Deep Learning 24
2.2.7 Class Imbalance 27
2.3 Conceptual Literature 30
2.3.1 Anomaly Detection 30
2.3.2 Deep Learning Architectures 33
2.3.3 Explainability and Interpretability Methods 38
2.3.4 Ensemble Learning Approaches 41
2.3.5 Class Imbalance Handling Techniques 44
2.4 Empirical Literature 46

2.4.1 Energy Consumption Anomaly Detection Studies 46
2.4.2 Empirical Comparison of Deep Learning Architectures..... 55
2.4.3 Performance Metrics for Evaluating Models 58
2.5 Ethical Considerations in Deep Learning 60
2.5.1 Privacy and Data Security..... 60
2.5.2 Fairness and Bias 60
2.5.3 Transparency and Accountability 61
2.5.4 Environmental and Social Impact 61
2.6 Literature Review Gaps 62
2.7 Conceptual Framework..... 63
2.7.1 Conceptual Elements 63
2.7.2 Operationalization of the Variables 65
2.8 Chapter Summary 66
CHAPTER THREE: METHODOLOGY..... 68
3.1 Introduction..... 68
3.2 Research Philosophy..... 68
3.3 Research Design..... 69
3.4 Experimental Design..... 71
3.4.1 Research Hypothesis..... 72
3.4.2 Pilot Study..... 72
3.4.3 Main Experiments..... 73
3.4.4 Integrated Development Environment & Libraries 75
3.4.5 Research Instruments..... 76
3.4.6 Research Population, Sampling and Sample Size..... 76
3.4.7 Data Collection Process and Techniques 80
3.4.8 Data Analysis Approach 81
3.5 Model Development..... 81
3.5.1 Application of Design Science Research in Deep Ensemble Model Development..... 82
3.5.2 Conceptual Design..... 84
3.5.3 Model Architecture..... 85
3.5.4 Model Design..... 86
3.5.5 Prototyping..... 87
3.5.6 Experiments 87
3.6 Ethical Considerations 88
3.7 Chapter Summary 88
CHAPTER FOUR: ANALYSIS OF MODELS FOR ANOMALY DETECTION BASED ON CLASS IMBALANCE..... 90
4.1 Introduction..... 90

4.2 Review of Existing Models 91
4.2.1.Traditional Anomaly Detection Methods 91
4.2.2.Deep Learning Models 92
4.2.3.Techniques for Handling Class Imbalance 93
4.3 Methodology 93
4.4 Results and Discussion 95
4.4.1 Classical Machine Learning Models 95
4.4.2 Deep Learning Models 96
4.4.3 Limitations of Existing Approaches 96
4.5 Key Insights 97
4.6 Discussion 97
4.7 Chapter Summary 100
CHAPTER FIVE: EMPIRICAL ANALYSIS OF CLASS IMBALANCE HANDLING MODELS 102
5.1 Introduction 102
5.2 Methodology 103
5.2.1 Data Collection and Preprocessing 103
5.2.2 Model Architecture and Implementation 104
5.2.3 Experimental Setup 105
5.2.4 Evaluation Criteria 106
5.3 Results of Empirical Analysis 107
5.3.1 Performance Metrics 107
5.3.2 Comparative Performance of Selected Models 107
5.4 Discussion 110
5.5 Chapter Summary 114
CHAPTER SIX: DEVELOPMENT OF DEEP ENSEMBLE LEARNING MODEL 116
6.1 Introduction 116
6.2 Architecture Design 117
6.2.1 Description of the Deep Ensemble Model 117
6.2.2 Components of the Architecture 118
6.2.3 Rationale for Design Choices 120
6.3 Implementation 121
6.3.1 Tools and Technologies 121
6.3.2 Data Preprocessing 122
6.3.3 Training and Validation Processes 122
6.4 Results of Model Development 127
6.5 Discussion 130
6.5.1 Comparison with Existing Literature 130

6.5.2 Challenges Faced During Development 131
6.5.3 Theoretical Underpinnings of the Developed Model..... 132
6.6 Chapter Summary 133
CHAPTER SEVEN: EVALUATION OF THE DEEP LEARNING ENSEMBLE MODEL 135
7.1 Introduction..... 135
7.2 Evaluation Methodology..... 135
7.2.1 Comparison with Baseline Models 136
7.2.2 Statistical Analysis..... 136
7.3 Results of Model Evaluation..... 137
7.3.1 Detailed Performance Results..... 137
7.3.2 Comparative Analysis with Baseline Models..... 138
7.3.3 Statistical Analysis..... 140
7.4 Discussion..... 145
7.4.1 Improvements in Model Design..... 146
7.4.2 Model Misclassifying Certain Anomalies 146
7.4.3 SHAP/LIME-Based Model Explainability 147
7.4.4 Deployment Trust 148
7.4.5 Real-World Deployment Feasibility 148
7.4.6 Model Error Mitigation Strategies 149
7.4.7 Implications of Findings 149
7.4.8 Limitations of the Current Study 150
7.5 Chapter Summary 151
CHAPTER EIGHT: SUMMARY, CONCLUSION AND RECOMMENDATIONS..... 153
8.1 Summary..... 153
8.2 Conclusion 157
8.3 Recommendations..... 159
8.4 Future Research 160
REFERENCES..... 163
APPENDICES 187

LIST OF TABLES

Table 3.1: Strategies to Achieve the Specific Objectives 71
Table 3.2: Experiments on the Various Models, Dataset and Class Imbalance Handling Techniques 74
Table 3.3: Python Libraries and Their Uses 75
Table 3.4: Description of the Experimental High-Dimensional Dataset 78
Table 3.5: Description of the Class Imbalance Handling Techniques 80
Table 5.1: Hyperparameter Configurations 106
Table 5.2: Baseline Models (Without Class Imbalance Handling) 108
Table 5.3: Baseline Models (With Class Imbalance Handling) 109
Table 6.1: BiLSTM Model Parameters 123
Table 6.2: CNN Model Parameters 124
Table 6.3: Data Preprocessing Parameters 125
Table 6.4: Ensemble Parameters 126
Table 6.5: Training and Validation Performance 127
Table 7.1: Performance Comparison of Models 137
Table 7.2: Results of Statistical Analysis 141
Table 7.3: Post-hoc Analysis using Tukey's HSD Test 143

LIST OF FIGURES

Figure 1.1: Class Imbalance Handling Techniques 3
Figure 2.1: Conceptual Framework 65
Figure 3.1: Conceptual Model Flowchart for Ensemble Model Development 92
Figure 3.2: Architecture of the Deep Learning Ensemble Model for Addressing the Class Imbalance Challenge 86
Figure 3.3: Research Methodology for System Development 88
Figure 6.1: The Final Deep Ensemble Model 119

ACRONYMS AND ABBREVIATIONS

AE	Autoencoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
BiLSTM	Bi-Directional Long Short-Term Memory
ConvLSTM	Convolutional Long Short-Term Memory
CNN	Convolutional Neural Networks
DNN	Deep Neural Network
EVT	Extreme Value Theory
FCN	Fully Connected Network
FPR	False Positive Rate
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
GPU	Graphics Processing Unit
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
ReLU	Rectified Linear Unit
RF	Random Forests
RNN	Recurrent Neural Network
SGCC	State Grid Corporation of China
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The landscape of energy consumption has undergone significant transformations recently with the integration of smart technologies such as smart meters and advanced metering infrastructure (AMI), resulting in more complex and interconnected systems[1]. Smart technologies are transforming how we monitor and control energy consumption. These technologies use sensors, data analytics, artificial intelligence (AI), and the Internet of Things (IoT) to deliver real-time insights and automated control over energy consumption. Smart meters serve as foundational components of modern energy systems, enabling real-time tracking of electricity consumption.

Monitoring and managing energy consumption are critical aspects of ensuring the efficiency, sustainability, and reliability of modern energy infrastructures[1],[2]. Anomaly detection in energy consumption datasets is pivotal in identifying irregularities, potential faults, or abnormal patterns that may indicate inefficiencies or security threats. However, the accurate detection of anomalies faces substantial challenges, mainly in the form of class imbalance within these datasets [2].Energy consumption datasets inherently exhibit class imbalance, where instances of normal energy consumption significantly outweigh instances of anomalous behaviour [3]. Class imbalance poses a significant challenge in energy consumption monitoring, particularly in anomaly detection systems for smart grids, because instances of "normal" energy usage overwhelmingly outnumber "anomalous" events like equipment faults, theft, or cyberattacks.

Machine learning models, when trained on such imbalanced datasets, tend to be heavily biased towards the majority class (normal behaviour), leading to a high rate of false negatives where critical anomalies are missed [4]. This occurs because traditional algorithms often optimize for overall accuracy, which can be misleading when the minority class (anomalies) is the one of paramount interest; a model might achieve 99% accuracy simply by classifying everything as "normal," thereby failing to detect the rare, yet crucial, deviations. Neural networks and classical classifiers typically assume balanced class distributions, leading them to “learn” features of the overrepresented class while underfitting or ignoring the minority class entirely. Recent

studies highlight that traditional resampling techniques (e.g., SMOTE) may introduce noise, while cost-sensitive learning and generative adversarial networks (GANs) show promise in improving detection rates[5]. Addressing imbalance is essential for reliable energy analytics, as undetected anomalies can escalate operational costs and grid instability.

Class imbalance can mislead model performance evaluation, especially when using accuracy as the primary metric. Imbalanced datasets often result in models leaning towards the majority class, yielding suboptimal performance for the minority class. Class imbalance can negatively affect a model's generalization ability. The traditional approaches to anomaly detection often struggle to adequately address this imbalance, leading to suboptimal performance in identifying and classifying anomalies. As a consequence, there is a critical need for advanced models that can effectively handle class imbalance, heighten the precision, recall, and overall accuracy of anomaly detection models [3]. Multiple approaches have been employed to tackle the class imbalance problem and improve model performance[6]. Addressing class imbalance in classification tasks can be achieved through various methodological approaches.

One prominent category involves data-level methods, which focus on manipulating the composition of the training dataset to mitigate the skewed class distribution. For example, oversampling the minority class endeavours to equalise class representation by increasing the cardinality of the under-represented class. This can be accomplished through techniques such as simple duplication of existing minority class instances or more sophisticated methods like the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE achieves its effect by generating synthetic minority class samples through a blend of pairs of current minority class instances in the feature space. Undersampling the majority class lower the number of cases of the majority class by randomly removing instances. The balance of classes is achieved in hybrid methods through the integration of oversampling and undersampling, such as SMOTE combined with Tomek links or Edited Nearest Neighbors (ENN).

Secondly, algorithm-level techniques deal with class imbalance by modifying the learning algorithm to mitigate a tendency for the model to favour the majority class. The cost-sensitive learning assigns differential misclassification penalties, with higher costs typically associated with errors on the minority class [7]. This approach can be

implemented by adapting algorithms including cost-sensitive decision trees or support vector machines (SVMs). Alternatively, class weighting techniques directly influence the learning process by giving greater importance to the minority class. For instance, in neural networks, class weights can be incorporated into the loss function to emphasize the contribution of minority class samples during training.

Thirdly, ensemble methods can be particularly effective for imbalanced datasets [8]. Ensemble learning techniques such as Boosting (e.g., Gradient Boosting) and Bagging (e.g., Random Forests) can be adapted to address class imbalance. For example, Balanced Random Forests balance the classes in each bootstrap sample. Balanced Random Forest (BRF) mitigates class imbalance by training each constituent tree on a balanced bootstrap sample. Figure 1.1 shows the popular class imbalance handling techniques

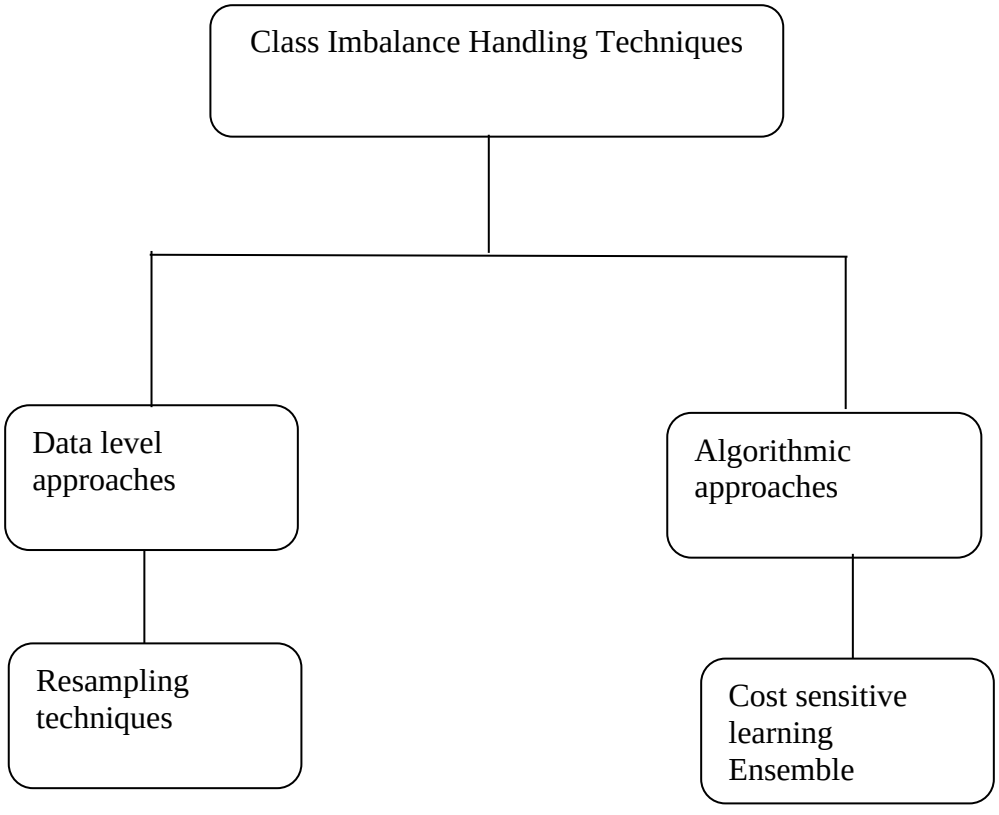


Figure 1.1: Class Imbalance Handling Techniques

Lastly, deep learning methods special techniques may be used to handle imbalanced data. Autoencoders for anomaly detection train autoencoders on the majority class (normal data) and use reconstruction error to detect anomalies (minority class) [9].

Generative Adversarial Networks (GANs) use GANs to synthesize synthetic minority class samples to balance the dataset [10].

Deep ensemble modelling has significantly enhanced anomaly detection and accuracy in various ways. Firstly, improved robustness and reliability by combining predictions from multiple models, deep ensembles mitigate the potential pitfalls of relying on one model that might be prone to specific errors or biases. This has led to more robust and reliable anomaly detection. If one model fails to identify an anomaly, others in the ensemble might still catch it, thus reducing false negatives and false positives. Secondly, uncertainty estimation deep ensembles provide an estimate of uncertainty in their predictions. This is predominantly useful in anomaly detection where understanding the confidence of the model's output can facilitate the distinction between normal and anomalous behaviour more accurately. Anomalies often lie in regions of high uncertainty, and ensemble models can capture this uncertainty better than single models.

Thirdly, enhanced accuracy ensembles often outperform individual models by leveraging the strengths of different models and averaging out their limitations. This generally leads to higher overall accuracy in anomaly detection. Fourthly, by fusing predictions from multiple models, deep ensembles tend to generalize better to new, out-of-sample data. This is crucial for anomaly detection, as anomalies are rare and diverse by nature. Ensembles reduce overfitting to training data and improve performance on varied datasets. Lastly, using different architectures, training procedures, or random initializations within an ensemble can increase model diversity. This diversity ensures that the ensemble covers a broader range of potential anomalies, improving detection capabilities. These benefits make deep ensembles a powerful tool in various domains requiring reliable anomaly detection.

Traditional solutions to class imbalance include resampling techniques such as oversampling (e.g., SMOTE) or undersampling [11]. While effective in some cases, these methods introduce their own limitations. Oversampling can lead to overfitting due to duplication or artificial data generation, while undersampling may result in loss of valuable information from the majority class. Moreover, these approaches manipulate the data distribution, which may not be desirable in time-series contexts where temporal dependencies are critical.

Cost-sensitive learning provides a principled alternative by modifying the learning process rather than the data. This technique assigns higher misclassification costs to the minority class, effectively guiding the model to pay more attention to the detection of anomalies during training[7]. By embedding the imbalance directly into the loss function, cost-sensitive methods allow for more robust learning without altering the original data distribution. This is especially valuable in energy systems, where preserving temporal structure and sequence integrity is crucial for model performance. Cost-sensitive learning requires providing different weights or costs to classes during the training process, with a focus on penalizing minority class misclassifications [12].

The adaptive cost-sensitive learning (ACL) method, which is effective with high imbalance ratios, dynamically adjusts the sample cost throughout the training process. The model is trained with weighted losses in a way that balances each class's contribution to model parameter updates, preventing the majority classes from dominating in imbalanced data during training. Sample costs are calculated by taking into account the sample number distribution, class convergence trend, and sample convergence trend [12]. This method is especially beneficial in applications like medical diagnosis and fraud detection, where precisely identifying under-represented groups is crucial [5]. Cost-sensitive learning was preferred in this study because it assigns differential misclassification penalties, with higher costs typically associated with errors on the minority class [7]. Cost-sensitive learning encourages the model to pay greater attention to correctly identifying instances from the minority class.

Machine learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forests, and Isolation Forests have been commonly used for anomaly detection in energy consumption data, particularly to address class imbalance. Standard machine learning algorithms, which often attempt to maximize overall accuracy, frequently fail to adequately learn the characteristics of the minority class due to underrepresentation [1]. This leads to models that are biased towards the majority class, resulting in a high number of false negatives for the minority class a critical issue when the penalty of missing an actual anomaly is high.

While various machine learning algorithms and techniques have been developed to tackle class imbalance, each comes with its own set of deficiencies, ranging from the risk of overfitting and information loss in sampling methods to the complexity and

tuning challenges of algorithm-level modifications and ensemble approaches[13],[5]. SVM struggles with scalability and kernel tuning in large datasets; KNN is sensitive to distance metrics and performs poorly in sparse regions; Random Forests may still favour majority classes; and Isolation Forest lacks temporal sensitivity. These deficiencies highlight the need for more advanced solutions, such as cost-sensitive deep learning models, that can better handle class imbalance in time series energy data.

1.2 Problem Statement

Anomalies in energy consumption refer to unusual or unexpected patterns and observations in the consumption of energy resources, deviating significantly from the usual normal [14]. Accurate detection of these anomalies is essential for energy conservation, cost reduction, and the identification of potential inefficiencies in energy systems [14]. There are multiple reasons for anomalies in energy consumption, including equipment failure, change of weather, energy theft and data collection errors from the sensor data streams [15]. Numerous machine learning and deep learning algorithms have been adapted to detect outliers in energy use. The "Anomaly Detection method of Smart Meters data based on GMM- LDA clustering feature Learning and PSO Support Vector Machine" is one of such machine learning models based on Support Vector Machines [16].

It uses binary classification to represent one class of normal data and another class of abnormal data, as well as the outliers [16]. However, Support Vector Machines struggle when trained on class imbalanced datasets, potentially fostering a bias that prefers the majority class and, as a result, decreasing anomaly detection accuracy [17], [18]. Energy consumption data often exhibits class imbalance, where normal data points significantly outnumber anomalies [19]. Besides this, tuning for the C regularization parameter and the choice of features in the Support Vector Machines is also challenging [19]. C-Regularisation parameter is used to control the strength of regularization to prevent overfitting. This is accomplished by incorporating a regularization cost in the cost function. Depending on the choice of these parameters and features, Support Vector Machines produce false positives or miss certain anomalies, impacting the overall accuracy of the energy consumption detection process [19]

K-Means is another common machine learning algorithm used to identify deviations in energy consumption [20]. It groups similar data points together, wherein the set of observations with anomalies is put in sparsely populated clusters [20]. The “Power consumption predicting and anomaly detection based on transformer and K-means” presents a model for detecting energy consumption anomalies using K-means. However, *k*-means struggle with class-imbalanced datasets where selecting the appropriate number of clusters, *k,* can be cumbersome, and a poor choice of the k-value leads to overestimation or underestimation of the clusters. Both have an impact on the level of accuracy in the detection of these anomalies [21].

Mao et al. [22] proposed the isolation forest to address the class imbalance problem. The isolation forest is a progressive ensemble method that isolates anomalies by building random decision trees and identifying data points with shorter average path lengths in the trees [22]. The “Anomaly detection for power consumption data based on isolated forest” is an example of an isolation forest-based model for detecting energy usage anomalies. However, tuning isolation forests' hyperparameters, including the number of trees and the sub-sampling size, can be challenging, and selecting the wrong hyperparameters makes it difficult to handle class-imbalanced datasets [23]. This subsequently affects the level of accuracy in anomaly detection in energy consumption.

Deep learning algorithms have also been employed to tackle class imbalance in anomaly detection for energy consumption. [24]. Autoencoders are one such algorithm that identifies anomalies through the use of the reconstruction error [25]. The “Detection of appliance-level abnormal energy consumption in buildings using autoencoders and micro-moments” is an example of an autoencoder-based model [25]. Choosing the right features as well as handling class-imbalanced datasets are the two main challenges with the autoencoder. Choosing the right features to include in the autoencoder model is essential because using irrelevant or redundant features often leads to suboptimal and inaccurate results [25]. A class imbalance problem in the autoencoders causes a bias towards the majority class and potentially missing rare anomalies, subsequently affecting the level of accuracy in anomaly detection.

Based on this, it is concluded that the lack of effective training strategies for mitigating class imbalance is one of the biggest challenges with existing models; class imbalance

leads to false positives and false negatives errors, which subsequently lead to decreased accuracies in the detection of anomalies.

1.3 Objectives

1.3.1 General Objective

This research aimed to develop and evaluate a deep ensemble model for anomaly detection in imbalanced energy consumption data.

1.3.2 Specific Objectives

The specific objectives of the research are:

i. To investigate existing models used to handle class imbalance in anomaly detection for energy consumption, with the aim of identifying effective techniques that can inform model selection for empirical analysis.
ii. To conduct an empirical analysis of class imbalance handling models in the domain of anomaly detection for energy consumption.
iii. To develop a deep ensemble model that integrates cost-sensitive learning, a class imbalance handling technique, for anomaly detection for energy consumption data.
iv. To evaluate the performance of the newly developed deep learning ensemble model in addressing the class imbalance challenge in the domain of anomaly detection for energy consumption.

1.4 Research Questions

i. Which models are used to address the class imbalance challenge in the domain of anomaly detection for energy consumption?
ii. How do the existing class imbalance handling models in the domain of anomaly detection for energy consumption compare in their performance?
iii. How can a deep ensemble model that integrates cost-sensitive learning technique for anomaly detection for energy consumption data be developed based on the insights from the theoretical and empirical analysis?
iv. How can the performance of the newly developed deep learning ensemble model be evaluated in the domain of anomaly detection for energy consumption?

1.5 Research Hypothesis

A research hypothesis is a proposed explanation for an existing phenomenon [26]. In this research, the null hypothesis was formulated as follows:

H_0: "There is no statistically significant difference in F1-score, accuracy, precision, recall, or AUC-ROC between a deep ensemble model without cost-sensitive learning and other deep learning-based anomaly detection models on imbalanced energy consumption data".

1.6 Significance of the Study

This study developed a novel model to deal with class imbalance in energy consumption anomaly detection, demonstrating applicability beyond academic contexts. At its core, the research holds the promise of transformative impacts on the resilience, efficiency, and security of contemporary energy systems. In the dynamic landscape of smart technologies and data-driven decision-making, the proposed model serves as a beacon, guiding the way to more accurate and reliable anomaly detection.

The prevalence of class imbalance within energy consumption datasets is a ubiquitous hurdle that conventional anomaly detection models struggle to surmount. By pioneering a specialized model tailored to the intricacies of the energy domain, this research intended to deliver a robust solution to address current limitations, with the potential to significantly advance anomaly detection in the energy sector, where anomalies can indicate inefficiencies, faults, or security threats.

The proposed model, equipped with the capacity to navigate class imbalance challenges, has the potency to act as an early-warning system, detecting irregularities that may have a considerable impact if left unaddressed. This not only fortifies the reliability of energy systems but also contributes to the overarching desired outcomes of sustainability and efficiency. Moreover, the impact of this research resonates in its practical applications. As energy systems become increasingly complex and interconnected, the need for precise anomaly detection mechanisms becomes non-negotiable. The proposed model, once validated and implemented, could find application in smart grids, energy management systems, and cybersecurity protocols.

Its pragmatic importance rests on its ability to enhance the decision-making processes

that govern the utilization of energy resources, ultimately contributing to the sustainable and efficient operation of energy infrastructures. On a broader scale, the significance of this research reverberates through the academic community, industry stakeholders, and policymakers alike. By advancing the frontier of anomaly identification in energy consumption datasets, this research has the potential to influence best practices, shape industry standards, and inform policy decisions. The study aligns with the global imperative to change towards smarter, more resilient energy systems that can meet the energy demands of the future. In summation, the impact of this research is its potential to usher in a paradigm shift in how anomalies are detected and addressed in energy consumption datasets. By navigating the complexities of class imbalance and offering a model attuned to the notable challenges of the energy domain, this research is evidence of the transformative power of advanced machine learning techniques. It is not merely an academic pursuit but a forward-looking endeavour with substantial ramifications for the reliability, efficiency, and security of our energy systems.

1.7 Scope of the Study

This research developed and evaluated a deep ensemble learning method for anomaly detection in imbalanced energy consumption data. The methodology incorporated cost-sensitive learning, to CNNs, and BiLSTM to mitigate class imbalance. The study focused specifically on deep learning approaches to enable in-depth analysis. Using a real-world energy consumption dataset provided by the State Grid Corporation of China (SGCC), the performance of the ensemble model was evaluated, where anomalous consumption was labeled as '1' and normal consumption as '0'. The study used accuracy, precision, recall, F1-score and AUC-ROC as evaluation metrics.

1.8 Limitations of the Study

Although the outcomes show that the ensemble model is highly effective, some limitations are noted:

i. **Dataset Dependency:** The model's performance was evaluated on a specific dataset (SGCC). The data was collected at one time, so it might not reflect future conditions.

ii. **Class Imbalance Handling**: Although the cost-sensitive technique effectively addressed the class imbalance in this study, exploring other advanced

techniques such as oversampling with SMOTE [27], under-sampling, or generative adversarial networks (GANs) [28] for handling imbalanced data could potentially further improve model performance.

iii. **Computational Cost:** Training deep learning models, especially complex ensemble models, can be computationally expensive. Exploring strategies like model compression and efficient training algorithms is critical for deploying these models in real-world applications.

iv. **Explainability:** Although the model achieved high performance, understanding the reasons behind its predictions can be difficult. Exploring techniques to enhance the model's explainability, such as attention mechanisms or feature importance analysis, is essential for building trust and helping users understand the results.

1.9 Contributions of the Thesis

This research has made several significant contributions across theoretical, practical, and policy domains:

Contribution to Theory

i. **Novel Deep Ensemble Model:** Developing a deep ensemble model integrating BiLSTM and CNN models within a heterogeneous framework constitutes a significant contribution to the anomaly detection field. This model validates the effectiveness of combining diverse deep learning models to boost performance and robustness.

ii. **Class Imbalance Handling in Deep Learning Ensembles:** The success in integration of cost-sensitive learning within the deep ensemble framework provides valuable insights into effective strategies for addressing class imbalance in deep learning models, particularly in the anomaly detection context.

Contribution to Practice

i. **Empirical Analysis of Class Imbalance Handling Techniques:** The empirical analysis of class imbalance handling techniques, including oversampling, undersampling, and cost-sensitive learning, offers valuable

insights into their efficacy for energy consumption anomaly detection. These findings underwrite a deeper understanding of the opportunities and challenges associated with dealing with class imbalance in this domain.

ii. **Enhanced Anomaly Detection Capabilities:** The developed deep ensemble model establishes improved accuracy of 97.5%, a precision of 97%, a recall of 99%, an F1-score of 98%, and an AUC-ROC score of 99% in comparison with individual base models and modest performance in comparison with the current best models from the literature. This improved performance has direct, tangible implications for improving the effectiveness of anomaly detection systems in real-world energy management applications.

iii. **Improved Grid Stability and Reliability:** Accurate and timely anomaly detection is central to maintaining grid stability and reliability. The developed model contributes to the early identification and mitigation of critical events such as equipment failures, sudden load changes, and cyberattacks, leading to a more robust and resilient energy grid.

iv. **Enhanced Energy Efficiency:** By accurately identifying and addressing inefficiencies and equipment malfunctions, the model contributes to significant energy savings, reducing operational costs and minimizing environmental impact.

1.10 Organization of the Thesis

Chapter One presents a brief introduction to the background of the study, problem statement, general and specific objectives, research questions, research hypothesis, significance of the study, research contributions and thesis outline.

Chapter Two presents a literature review of the existing models that deal with class imbalance challenges in the realm of anomaly detection for energy consumption. A theoretical framework was discussed, followed by conceptual literature on anomaly detection, deep learning architectures, explainability and interpretability methods, ensemble learning approaches, class imbalance handling techniques, empirical literature, performance metrics for evaluating models, ethical considerations, literature review gaps and conceptual framework. The identification of gaps from the literature review and the key findings culminates this chapter.

Chapter Three presents the methodology of this work. The culmination of this analysis

was a conceptual model that subsequently guided the empirical research design. The experimental results formed the basis for the formulation of the deep ensemble model. This development was guided by a design science research design. In the end, the approach for evaluating the efficacy of the newly developed model was evaluated against the existing ones using the same dataset comprising many variables and the measures used to determine model success.

Chapter Four presents the analysis of existing anomaly detection models for energy consumption data. This chapter focuses on the challenges posed by class imbalance and explores existing solutions.

Chapter Five presents the results of the empirical analysis and performance of various deep learning models for anomaly detection in energy consumption data, specifically focusing on the impact of class imbalance handling techniques.

Chapter Six details the development of the proposed deep ensemble learning model, including the selection of base models (BiLSTM and CNN), the integration of class imbalance handling technique, cost-sensitive learning, and the design of the ensemble architecture.

Chapter Seven reports the findings of the model evaluation, comparing the performance of the proposed ensemble model with individual base models and current best models from the literature, including those proposed by Huang et al. and Gao et al.

Chapter Eight summarizes the key findings of the research, discusses the meaning of the findings within the field of anomaly detection in energy consumption data, and puts forward recommendations for future inquiry.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The increasing use of smart technology in energy consumption systems has transformed the way electricity is monitored and managed [29]. The growing prevalence of smart grids and advanced metering infrastructure generates substantial data, offering opportunities for enhanced anomaly detection in energy consumption[30]. However, a substantial challenge arises from the inherent class imbalance in anomaly data, where normal instances immensely outnumber abnormal ones[31]. The class imbalance issue poses a substantial hurdle in the development of effective anomaly detection models [31]. When faced with imbalanced datasets, traditional machine learning algorithms often exhibit a preference for the majority class, resulting in inferior performance in detecting uncommon anomalies that are crucial for energy system stability and efficiency [29]. In recent years, addressing class imbalance has become a focal point of research, with a collection of models and techniques proposed to improve the robustness of anomaly identification in energy usage data.

This review critically synthesizes the literature on models designed for class imbalance in energy anomaly detection. It elucidates the strengths, limitations, and potential advancements of current approaches applied to imbalanced energy consumption datasets.

This review examined various models designed to tackle class imbalance in energy consumption anomaly detection, evaluating their algorithmic adaptability, scalability, and deployment potential. It further analysed how transfer learning, ensemble methods, and novel loss functions have been employed to enhance anomaly detection accuracy in the presence of class imbalance.

2.2 Theoretical Review

This section examines and synthesises key theories, models, and conceptual frameworks relevant to this study.

2.2.1 Computational Theory

Computational theory has evolved into a vital domain in machine learning and data analytics, particularly in the context of energy consumption analysis. Recent advancements have underscored the complexity of identifying irregular patterns in large-scale energy datasets, where traditional detection methods often weaken as a result of inherent data imbalances [32], [33]. The fundamental premise of computational anomaly detection lies in developing sophisticated algorithms capable of distinguishing between normal patterns and statistically significant deviations that may indicate potential failures, systemic inefficiencies, or extraordinary consumption events.

In domains characterized by highly skewed datasets, ensemble learning techniques have gained substantial traction in addressing the challenges of anomaly detection. Ahmed, et al.[34] demonstrated that deep ensemble models can substantially improve detection accuracy by leveraging multiple learning paradigms, effectively mitigating the limitations of individual algorithmic approaches. The researchers proposed the Deep Ensemble Learning Model (DELM), an effective anomaly detection system for malicious network traffic attacks. Their research underscores the potential of combining diverse neural network architectures to enhance anomaly identification capabilities, particularly in network intrusions scenarios where traditional statistical methods prove inadequate. To mitigate the challenge of limited data for minority classes, a Conditional Generative Network was integrated to enhance DELM's data generation capabilities. Model accuracy was further improved through preprocessing steps like data augmentation and batch normalization. For effective feature extraction, they employed n-gram, feature aggregation, and one-hot encoding techniques.

The theoretical foundations of anomaly detection are increasingly anchored in advanced deep learning methodologies, with particular emphasis on generative and discriminative models. A data generative model (DGM) employing a Conditional Generative Adversarial Network (CGAN) was proposed [35] to augment minority classes within the Intrusion Detection System (IDS) dataset, thereby addressing inherent challenges associated with imbalanced datasets. Their work highlights the importance of sophisticated representation learning for distinguishing rare anomalies within a dataset dominated by normal instances. These computational techniques

provide more adaptive and context-aware anomaly detection compared to traditional statistical methods.

A critical evaluation of current computational theories reveals significant limitations in existing anomaly detection frameworks. While deep learning models demonstrate remarkable performance, they often struggle with interpretability and computational complexity[36],[37]. The inherent lack of transparency in many ensemble models presents significant challenges to the interpretability of their decision-making processes, particularly within critical domains such as energy management. Furthermore, the computational burden associated with complex ensemble architectures can potentially constrain their applicability in real-time scenarios. Subsequently, there is a pressing need for innovative methodologies that find a trade-off between detection accuracy and computational efficiency.

The proposed research objective, namely the development of a deep ensemble model for anomaly detection in imbalanced energy consumption data, directly addresses several critical gaps within contemporary computational theory. The incorporation of multiple deep learning architectures within the proposed model seeks to mitigate the inherent issues associated with single-model approaches. Preliminary studies by Liu, et. al. [38] suggest that ensemble methods can improve anomaly detection accuracy in complex energy datasets, providing a compelling rationale for the proposed research methodology. The researchers proposed SEAD-PL, 'Selective Ensemble method for Anomaly Detection based on Parallel Learning' to reduce training time and resource usage, a distributed parallel training framework. Subsequently, a clustering-based ensemble selection strategy serves to balance the accuracy and diversity of the constituent base detectors. The approach seeks to develop a flexible, adaptive framework capable of dynamically adjusting to varying data distributions and capturing subtle consumption irregularities.

In their work, [39] proposed a method for detecting anomalies in smart home energy consumption data. The framework uses a machine learning-based prediction model to forecast energy use and then identifies any deviations from the predicted values as anomalous consumption. This framework proposes a two-stage approach to smart home energy anomaly detection. First, an anomaly detection module employing local outlier factor (LOF), connectivity-based outlier factor (COF), and cluster-based local

outlier factor (CLOF) identifies anomalous energy consumption. Second, a predictive model forecasts typical energy patterns using linear regression, decision tree, random forest, and ensemble boosting methods (gradient, adaptive, and extreme gradient) trained on anomaly-free data. Evaluated on three public smart home energy datasets, the framework demonstrates promising results, with CLOF coupled with extreme gradient boosting achieving the highest prediction accuracy. Experimental results demonstrated the strong performance of cluster-based local outlier factor (CLOF) combined with extreme gradient boosting, achieving high prediction accuracy.

Practical applications of computational anomaly detection theory extend beyond theoretical constructs, presenting profound effects on energy and sustainability. Kardi, et. al. [14] presented a novel, two-stage, deep learning approach for one-hour-ahead anomaly detection in electricity consumption data. First, an LSTM-based neural network predicts the subsequent hour's consumption. Second, an LSTM autoencoder, trained on normal consumption data (using the first stage's output as input), learns representative features. Anomalous consumption is identified by significant differences between the autoencoder's input and output. An Exponential Moving Average (EMA) threshold distinguishes local and global anomalies. The model incorporates meteorological (cloud cover, pressure, wind direction, humidity, wind speed, and temperature), temporal, and lag features. Feature selection is employed to optimize model performance achieving better results.

Fadlia, et al [40] research examined deep learning's potential in sports facility anomaly detection, overcoming traditional method constraints. By applying Deep Feedforward Neural Networks (DFNNs) to feature extraction and threshold estimation, the researchers developed an approach that minimizes false alarms. Validation using a Qatar University aquatic center dataset revealed the method's effectiveness, achieving 94.33% accuracy and a 92.92% F1-score.

In the domain of anomaly detection in class-imbalanced energy consumption datasets, computational theory provides a conceptual foundation for defining and designing learning algorithms capable of detecting rare or aberrant energy usage patterns[36],[37]. This theoretical foundation assures that the system is not only functionally efficient but also interpretable and adaptive to real-world settings in unbalanced domains such as energy analytics [36]. The current research shows that

computational models can effectively detect energy inefficiencies, enabling proactive conservation strategies. Our proposed ensemble model advances this field by identifying anomalies and uncovering the mechanisms behind irregular energy consumption, thus connecting computational insights with practical energy management approaches.

2.2.2 Extreme Value Theory

Extreme Value Theory (EVT) is a statistical theory for dealing with extreme deviations from the median of a distribution [41]. While it is often applied in finance, meteorology, and environmental science to model extreme events, it can also be leveraged in anomaly identification [41]. EVT is especially useful for detecting anomalies when anomalies are associated with rare and extreme occurrences in a dataset [42]. Extreme Value Theory can be applied to anomaly detection in modeling extreme events, threshold selection, block maxima or peak over the threshold, tail modeling, anomaly score calculation, dynamic thresholds, risk assessment, handling skewed distributions, and applications in time series data [43]. EVT focuses on modeling the tails of a distribution, specifically the extreme values [43]. In anomaly detection, these extreme values may correspond to unusual patterns or rare events that deviate significantly from the norm [41], [43]. By modeling extreme events, EVT allows one to set appropriate thresholds for anomaly detection. Rather than using a fixed threshold, EVT helps identify values in the tail of the distribution as anomalies[44], [41]. This is especially relevant when anomalies are rare but have a significant impact [44]. EVT employs two common approaches: Block Maxima and Peak Over Threshold [44]. Block Maxima entails separating the data into non-overlapping blocks and selecting the highest value from each. Peak Over Threshold focuses on values that exceed a specific threshold. Both methods are applicable in identifying extreme events.

EVT often involves fitting a distribution to the tail of the data [44]. Popular distributions include the Generalized Pareto Distribution (GPD) for Peak Over Threshold methods and the Generalized Extreme Value (GEV) distribution for Block Maxima methods. These distributions help estimate extreme event probabilities. Anomalies can be identified by calculating an anomaly score based on the likelihood of observing a given value according to the modeled distribution [43]. The higher the

anomaly score, the more likely the value is an anomaly. EVT allows for the establishment of dynamic thresholds that is adapted to the changes in data distribution. This is especially advantageous in situations where characteristics of normal and anomalous behaviour may evolve over time.

EVT is well-suited for analysing time series data, especially when anomalies are associated with extreme values in the temporal dimension [42],[43]. It enables the identification of periods with significantly deviant behaviour. Anomaly detection often encounters skewed distributions where most values are concentrated around a mean. EVT provides a robust methodology for identifying anomalies in the tails of such distributions. EVT not only helps in anomaly detection but also facilitates risk assessment by providing insights into the likelihood of extreme events. This can be valuable for proactive risk management in various domains [44],[43]. While EVT offers powerful tools for anomaly detection, its application necessitates careful assessment of the underlying assumptions and data characteristics. Additionally, the success of EVT in anomaly detection depends on the appropriateness of the chosen distribution and the quality of the modelling process [41].

2.2.3 Systems Theory

Systems theory accords a robust framework for analysing the intricate dynamics of complex systems, making it particularly appropriate to the proposed research on a deep ensemble model for anomaly detection in imbalanced energy consumption data. At its core, systems theory emphasizes the interconnectedness of components and the emergent behaviour arising from their interactions [45],[46] . This perspective is crucial for understanding energy systems, which can be conceptualized as complex adaptive systems (CAS).

Energy systems, encompassing generation, transmission, distribution, and consumption, exhibit characteristics of CAS, including adaptability, non-linearity, and emergent properties [47]. These systems are constantly adapting to changing environmental conditions, fluctuating demand, and technological advancements. Applying a systems theory lens allows us to move beyond a reductionist view, recognizing that energy consumption patterns are not merely the sum of individual behaviours but are shaped by complex interactions within the system.

Furthermore, feedback mechanisms are integral to the stability and performance of energy systems [48]. In the context of anomaly detection, real-time feedback on energy consumption patterns can trigger corrective actions, preventing cascading failures and maintaining system stability. For example, sudden spikes in energy consumption can be identified as anomalies, prompting automated responses to regulate power flow or alert operators to potential issues. By incorporating feedback loops into the deep ensemble model, we can extend its capacity to detect and respond to anomalies in a timely manner. The model can learn from its own predictions and adapt its parameters to improve accuracy and reduce false positives, thereby ensuring system stability.

Resilience and robustness are critical attributes of modern energy infrastructure [49]. Energy systems must be able to withstand disruptions, such as equipment failures, cyberattacks, and extreme weather events, while maintaining essential services. Systems theory affords a framework for analysing the resilience of energy systems by considering the interconnectedness of components and the capacity of the system to adapt to disturbances. In the proposed research, the deep ensemble model can contribute to enhanced resilience by identifying anomalies that may indicate impending failures or vulnerabilities. By detecting and mitigating these anomalies, we can strengthen the robustness of the energy infrastructure and minimize the impact of disruptions.

The concept of emergence and self-organization is also central to systems theory [50] In energy consumption patterns, emergent behaviour refers to the collective patterns that arise from the interactions of individual consumers and devices. For instance, peak demand periods are emergent phenomena resulting from the synchronized behaviour of many consumers. Self-organization refers to the capacity of the system to adapt and reorganize in response to changes in its environment. By applying a deep ensemble model, we can capture these emergent and self-organizing patterns, enabling the detection of anomalies that may not be apparent at the individual component level. The model can learn the underlying structure of normal energy consumption patterns and identify deviations that indicate anomalies.

Moreover, the challenge of imbalanced data, inherent in energy anomaly detection where normal consumption vastly outweighs anomalous events, further necessitates a

systems-level approach. The deep ensemble model, when grounded in systems theory, can effectively manage this imbalance by capturing the intricate relationships within the data. This is achieved through the ensemble's ability to combine multiple models, each potentially focusing on different aspects of the system's behaviour, thereby improving anomaly detection accuracy even in the presence of significant data skew.

In summary, systems theory provides a powerful framework for understanding and analysing the complex dynamics of energy consumption systems. By considering energy systems as CAS, incorporating feedback mechanisms, emphasizing resilience and robustness, and recognizing emergent behaviour, the proposed deep ensemble model can contribute to more effective anomaly detection and improved energy system performance.

2.2.4 Statistical Learning Theory

Statistical learning theory offers a rigorous mathematical framework for analysing the capabilities and limitations of machine learning algorithms, including deep ensemble models. Its principles are particularly relevant to the proposed research on anomaly detection in imbalanced energy consumption data, offering insights into model generalization, complexity, and performance. A cornerstone of statistical learning theory is the Probably Approximately Correct (PAC) learning framework. This framework aims to quantify the probability that a learned model will generalize well to unseen data [51]. In the context of energy consumption anomalies, PAC learning can help assess the likelihood that a deep ensemble model, trained on historical data, will accurately detect anomalies in future consumption patterns. This is vital, especially given the dynamic nature of energy consumption and the potential for shifts in underlying distributions.

The Vapnik-Chervonenkis (VC) dimension is another crucial concept, measuring the complexity of a model's hypothesis space [52]. A higher VC dimension indicates a greater capacity to fit complex patterns in the training data, further elevating the potential for overfitting. In the proposed research, understanding the VC dimension of the deep ensemble model is essential for balancing model complexity with generalization performance. By controlling the complexity of individual models within the ensemble, and the ensemble structure itself, we can mitigate overfitting and improve anomaly detection accuracy.

The bias-variance trade-off is a fundamental principle in statistical learning, and it is particularly relevant to ensemble methods [53]. Bias represents the error arising results from approximating a real-world problem using a simplified model, whereas variance indicates the model's sensitivity to variations in the training data. Ensemble methods mitigate variance and enhance overall prediction accuracy by aggregating multiple models. However, careful consideration must be given to the bias of individual models and the diversity of the ensemble to achieve optimal performance. In relation to imbalanced energy data, this means that individual models within the ensemble should exhibit sufficient diversity to capture different facets of the data, while also maintaining a low bias to properly model the anomaly events.

Empirical risk minimization (ERM) is a core principle guiding the training of machine learning models [54]. ERM seeks to minimize the error of the model on the training data. Nevertheless, in anomaly detection with imbalanced data, directly minimizing empirical risk can result in biased models that favour the majority class (normal consumption). Therefore, adaptations of ERM, such as cost-sensitive learning or resampling techniques, are necessary to address the class imbalance and ensure effective anomaly detection learning by the model. Furthermore, when applying deep learning to time series energy data, the temporal dependencies within the data add complexity to the ERM framework.

In the recent years, research has focused on adapting statistical learning theory to deep learning models, particularly in the context of complex datasets and anomaly detection. For example, researchers have explored methods for estimating the VC dimension of deep neural networks and developing regularization methods to enhance generalizability[55]. Furthermore, advances in understanding the generalization properties of ensemble methods have led to the development of more robust and effective anomaly detection algorithms.

In Summary, Statistical learning theory provides the mathematical backing needed to create a robust Deep Ensemble Model. By carefully considering the PAC learning framework, VC dimension, bias-variance trade-off, and empirical risk minimization principles, we can develop a model that effectively detects anomalies in imbalanced energy consumption data.

2.2.5 Traditional Machine Learning Algorithms

Class imbalance poses a significant challenge in energy consumption anomaly detection, where abnormal usage patterns often constitute a small minority of the dataset. Traditional machine learning algorithms such as Decision Trees, Support Vector Machines (SVMs), Random Forests, and Logistic Regression have been widely applied to this problem. However, these algorithms are inherently biased toward the majority class, especially when trained on skewed datasets. As a result, they tend to produce high overall accuracy but fail to adequately identify rare but critical anomalies. In the energy domain, where false negatives can lead to operational inefficiencies or security vulnerabilities, addressing this limitation is essential.

Decision tree classifiers partition the feature space recursively using attribute values that maximize a specific criterion. For this method, each node in the tree represents a dataset input value. The Decision Tree is created by separating the training dataset into subsets that only contain data from a single class. Decision trees are prone to bias towards the majority class because splitting criteria may favour the dominant class. Solutions to this problem include the use of class weights or pruning techniques that retain minority class performance[56] .

Random forest is a supervised machine learning framework that can handle both regression and classification tasks [57]. Random Forests are ensembles of decision trees trained on different bootstrap samples, with random feature selection at each split. RF underperforms in minority classes due to ensemble voting bias. To make predictions, RF uses majority voting. The solution to class imbalance is to utilise class weighting or balanced random forest, with each tree trained on a balanced bootstrap (equal minority and majority samples) [58],[57].

K-Nearest Neighbours (KNN) classifies an instance using the majority class of its k nearest neighbours in feature space. KNN Minority samples generally perform poorly when surrounded by majority class samples[20].

Support Vector Machines (SVMs) maximize the margin between classes by utilizing support vectors. The hyperplane is skewed towards the majority class. Solutions use class weight adjustment in the cost function or cost-sensitive SVMs that penalize misclassification of the minority class more [18]. Logistic regression models binary

outcomes by estimating the probability of a record belonging to a relevant class. The response variable has only two possible outcomes, represented by a binary indicator variable with values of 0 and 1. Logistic regression models the probability of a class using a logistic (sigmoid) function. Without regularization or class weights, logistic regression leans toward the majority class [59].

2.2.6 Deep Learning

Deep learning, a subset of machine learning, revolves around the intricate architecture of neural networks with numerous layers, known as deep neural networks [60]. At its core, deep learning fundamentally transforms data through hierarchical representations, unlocking its potential for tasks such as natural language processing, image recognition, and speech understanding [61]. Neural networks serve as the foundational building blocks of deep learning algorithms. These networks are made up of interconnected nodes organized into layers that include: an input layer, one or more hidden layers, and an output layer [62]. The essence lies in the activation functions, such as the rectified linear unit (ReLU) or sigmoid embedded within each node, introducing non-linearities that enable the model to learn complex relationships in data [60], [63].

Training deep neural networks involves a process known as backpropagation. The optimization algorithm minimizes the discrepancy between predicted and actual results by adjusting the network's weights and biases. The concept of gradient descent plays a pivotal role, in gradually fine-tuning the model through successive iterations [61],[64]. Encompassing a spectrum of architectures, deep learning boasts specialized models tailored for diverse data types.

Convolutional Neural Networks (CNNs) excel at image processing by using convolutional layers to extract hierarchical features. Recurrent Neural Networks (RNNs) can handle sequential data while keeping hidden states to capture temporal dependencies, making them ideal for tasks such as natural language processing [64],[14]. Autoencoders represent unsupervised learning models designed for feature learning and dimensionality reduction [25]. Autoencoders, comprising an encoder and a decoder, reconstruct input data and find applications in data compression and anomaly detection. Generative Adversarial Networks (GANs) employ a generator-discriminator framework to synthesize novel data samples, contributing to tasks such

as image generation and style transfer [62], [10]. The advent of transfer learning enhances deep learning's efficiency by leveraging knowledge from pre-trained models [24]. This practice involves fine-tuning models on new tasks, saving valuable computational resources and time.

Hyperparameter tuning becomes an essential aspect, influencing factors like learning rates, network architecture, and regularization techniques[24]. Deep learning frameworks like TensorFlow, PyTorch, and Keras make it easier to build and train these complex models [65]. These frameworks provide tools and abstractions, streamlining the development process for practitioners and researchers [65]. The interpretability of deep learning models poses a challenge due to their complexity.

Trade-offs often emerge between model interpretability and performance, prompting researchers to explore explainability techniques like feature importance analysis to shed light on the model's decision-making process [65], [66]. In conclusion, the fundamentals of deep learning algorithms lie in the transformative power of neural networks, their capacity for hierarchical representation learning, and the versatility of specialized architectures for diverse data types. As the field evolves, the ongoing interplay between innovation, interpretability, and practical applicability continues to shape the landscape of deep learning [65].

The universal approximation theorem, demonstrating that a feedforward neural network with a single hidden layer can approximate any continuous function, provides a theoretical basis for deep learning's power [67] However, recent research emphasizes that while this theorem establishes the possibility of approximation, it does not guarantee efficient learning or generalization, particularly in high-dimensional and imbalanced datasets [68],[69]. For energy consumption data, the theorem suggests that a sufficiently complex deep learning model can, in principle, capture the intricate patterns of normal and anomalous behavior. However, the practical challenge lies in designing and training such a model effectively, given the potential for overfitting and the need to handle imbalanced classes. Contemporary studies have explored the impact of network depth and width on approximation capabilities, highlighting that deeper networks may offer advantages in learning highly complex functions, but also increasing the risk of vanishing gradients and overfitting [70] Thus, while the universal approximation theorem validates the potential of deep learning, it necessitates careful

architectural design and training strategies.

The information bottleneck principle offers insights into the trade-off between compression and prediction in deep learning [71]. It suggests that an optimal representation of the input data should retain only the information relevant to the target task, while discarding irrelevant details. In the context of anomaly detection, the information bottleneck can guide the design of deep learning models that effectively extract features indicative of anomalies, while filtering out noise and irrelevant fluctuations in energy consumption. Recent works have extended this principle to understand the generalization capabilities of deep neural networks, demonstrating that models that achieve a better information bottleneck tend to generalize better [72]. However, implementing the information bottleneck in practice requires careful consideration of the information measures used and the optimization techniques employed. The challenge is to find a balance between compressing the input data and retaining sufficient information for accurate anomaly detection, especially when dealing with the subtle variations that may characterize anomalous events.

Regularization techniques contribute significantly in preventing overfitting and improving the generalization performance of deep learning models [73]. In the context of imbalanced energy consumption data, regularization is essential for ensuring that the model does not simply memorize the majority class (normal consumption) but learns to generalize to the minority class (anomalies). Common regularization methods, such as L1 and L2 regularization, dropout, and batch normalization, have been expansively studied. Recent research has directed at developing adaptive regularization techniques that can dynamically adjust the regularization strength based on the characteristics of the data and the learning process [74]. For example, multiple studies have looked into the use of adversarial training as a form of regularization, which can enhance the robustness of deep learning models to adversarial examples and improve their generalization performance.

Moreover, in time series anomaly detection, the temporal dependencies within the data necessitate the use of regularization techniques that can capture these dependencies while preventing overfitting, such as recurrent regularization or temporal dropout [75]. In summary, while regularizations are a necessity for deep learning, they must be chosen and tuned carefully to avoid underfitting, while still avoiding

overfitting. Tackling the issues associated with imbalanced data and complex temporal dependencies requires a nuanced understanding of these principles and the development of tailored deep learning architectures and training strategies.

2.2.7 Class Imbalance

Anomaly detection, a crucial facet of data analysis, is tasked with identifying patterns that deviate significantly from the norm [60]. In the realm of anomaly detection, one formidable challenge that researchers and practitioners grapple with is the class imbalance inherent in real-world datasets, for example, the energy consumption datasets [76]. This challenge becomes particularly pronounced in scenarios where anomalies are rare events, outnumbered by the more prevalent instances of normal behaviour [77]. In this section, the nuanced landscapes of class imbalance challenges in anomaly detection are explored, delving into the implications for model performance and the innovative solutions developed to address these challenges.

Class imbalance manifests when the distribution of instances in a dataset across distinct classes is unbalanced [30]. With respect to anomaly detection, anomalies typically represent the minority class, making up only a small fraction of the overall data. Traditional machine learning models, when trained on imbalanced datasets, favour the majority class, as their primary objective is to minimize overall classification error [30]. This poses a significant hurdle in anomaly detection, as the algorithm may struggle to discern and accurately classify rare instances of abnormal behaviour. The consequences of class imbalance in anomaly detection are multifaceted. One major implication is a considerable reduction in the sensitivity or recall of the model, leading to an increased likelihood of false negatives [29]. In circumstances where the cost of missing true anomalies is high, such as in detecting faults in critical systems or identifying fraudulent activities, this deficiency can have severe consequences. The challenge is compounded by the fact that anomalies often carry higher significance and require specialized attention [77].

To extenuate the impact of class imbalance, researchers have explored a variety of innovative approaches. Resampling techniques, such as oversampling the minority class or undersampling the majority class, geared to balancing the class distribution and make the learning algorithm more attuned to rare instances [61],[78].

Synthetic Minority Over-sampling Technique (SMOTE) is a notable example that generates synthetic instances of the minority class, effectively expanding its representation in the dataset [76]. Cost-sensitive learning is another strategy that involves assigning different misclassification costs to different classes [7]. By penalizing errors on the minority class more heavily, models are compelled to allocate greater attention to the rare anomalies [7]. Ensemble methods, like as Random Forests or Gradient Boosting, have proved effectiveness in handling class imbalance by combining the outputs of multiple base learners [77].

In conclusion, class imbalance in anomaly detection is a pervasive challenge with far-reaching implications for model performance and decision-making in critical applications. However, the dynamic landscape of research and innovation continues to yield strategies and methodologies that push the boundaries of what is achievable. The quest to refine anomaly detection techniques in the face of class imbalance reflects a commitment to enhancing the reliability and precision of models, ensuring their efficacy in safeguarding systems, detecting irregularities, and bolstering the overall robustness of anomaly detection in diverse and complex datasets.

The imbalance ratio, defined as the ratio of majority to minority class samples, significantly impacts the performance of learning algorithms [79]. High imbalance ratios can lead to biased models that predominantly classify samples as belonging to the majority class, giving rise to poor detection rates for the minority class (anomalies). In energy consumption data, where anomalies are inherently rare, the imbalance ratio can be extremely high, exacerbating this problem. Recent studies have demonstrated that the severity of the imbalance effect varies depending on the learning algorithm and the characteristics of the data [80].

Deep learning models, while powerful, are not immune to this issue. They can become overly specialized in recognizing the majority class, failing to capture the subtle patterns indicative of anomalies. Research has investigated several strategies to address class imbalance, including resampling techniques (e.g., minority oversampling, majority undersampling), cost-sensitive learning (differential misclassification costs for minority instances), and ensemble methods (aggregation of models trained on varied data subsets)[81]. However, the choice of technique and its effectiveness are contingent on the dataset and algorithm employed. Furthermore, with

the growing size of energy data, resampling techniques may become computationally expensive.

Theoretical bounds on minority class detection provide implications into the fundamental limits of learning algorithms in imbalanced datasets. These bounds often relate to the complexity of the minority class distribution and the degree of overlap between the majority and minority classes [82]. For example, studies have derived theoretical bounds on the error rate of minority class detection as a function of the imbalance ratio and the VC dimension of the learning algorithm. These bounds can help assess the feasibility of achieving a desired level of detection accuracy and guide the selection of appropriate learning algorithms and regularization techniques.

Recent research has also delved into use of information-theoretic measures to quantify the difficulty of minority class detection, providing a more subtle understanding of the factors that contribute to the imbalance problem [83],[84]. However, establishing tight theoretical bounds for complex deep learning models remains a challenge. The non-convex nature of deep learning optimization and the high dimensionality of the feature space make it difficult to derive analytical expressions for the error rate and generalization performance.

In the context of deep ensemble models, the challenge of imbalance is compounded by the need to ensure diversity among the ensemble members. If all models are trained on the same imbalanced dataset, they may exhibit similar biases, reducing the effectiveness of the ensemble. Recent studies have investigated strategies for creating diverse ensembles in imbalanced datasets, such as using different resampling techniques for each ensemble member or training models on different feature subsets[85]. Furthermore, the selection of evaluation metrics is critical for imbalanced datasets. This is due to the fact that a high accuracy can be achieved by simply predicting the majority class dominance. Precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) offer a more accurate evaluation of minority class detection performance [86]. It is important to remember, that in a very imbalanced problem, even a high AUC-ROC may not be enough, and the focus may need to shift to precision and recall.

To sum up, class imbalance poses a significant challenge to anomaly detection in

energy consumption data. Understanding the effects of imbalance ratios on learning algorithms and the theoretical bounds on minority class detection is essential for developing effective deep ensemble models. The development of adaptive learning algorithms should be a key focus of future research that can contextually adapt to varying imbalance ratios and on deriving tighter theoretical bounds for deep learning models in imbalanced datasets.

2.3 Conceptual Literature

Conceptual literature explores, defines, and develops concepts, ideas, and theoretical constructs related to this study.

2.3.1 Anomaly Detection

Anomaly detection in energy consumption is a critical task in modern energy systems, aimed at identifying irregular patterns that deviate from normal usage behaviour. These anomalies may indicate faults in metering infrastructure, equipment malfunctions, energy theft, or unexpected demand surges, all of which can compromise grid stability, operational efficiency, and economic performance. With the widespread deployment of smart meters and advanced metering infrastructure (AMI), utilities now collect high-frequency, time-stamped energy consumption data at the household and grid levels. This abundance of data presents both opportunities and challenges for automated anomaly detection.

Traditional rule-based methods and statistical approaches often struggle to scale or adapt to complex, dynamic usage behaviours. As a result, machine learning and deep learning techniques have gained prominence due to their ability to learn non-linear patterns, model temporal dependencies, and generalize from historical data. However, effective anomaly detection remains challenging due to factors such as data noise, seasonal variation, and the high class imbalance between normal and anomalous instances. These challenges necessitate the development of intelligent and adaptive models capable of detecting rare but impactful anomalies in real time, making anomaly detection an essential component of smart energy analytics and sustainable energy management.

2.3.1.1 Taxonomy of Anomaly Types in Energy Data

Energy consumption data exhibits a wide range of anomalies, reflecting the diverse

nature of operational disruptions and malicious activities. A robust taxonomy is essential for developing effective detection algorithms. Recent research has emphasized the need for a multi-faceted classification, considering both the source and the impact of anomalies [87]. For example, anomalies can stem from equipment failures (e.g., transformer faults, sensor malfunctions), cyberattacks (e.g., denial-of-service, false data injection), or operational errors (e.g., human mistakes, misconfigured settings).

Furthermore, the impact of anomalies can vary, ranging from localized disruptions to widespread blackouts. In the context of smart grids, the increasing integration of distributed energy resources (DERs) and advanced metering infrastructure (AMI) has introduced new anomaly types, such as voltage fluctuations caused by intermittent renewable energy generation and data integrity issues arising from compromised AMI devices [88]. Hence, a dynamic and adaptable taxonomy is required to capture the evolving landscape of energy anomalies.

Anomalies can be further characterized as contextual, point, or collective. Point anomalies refer to individual data points that vary considerably from the normal behaviour, for example sudden spikes or drops in energy consumption. Contextual anomalies, then again, are data points that are anomalous only in particular instances, such as a low energy consumption during peak demand hours. Collective anomalies involve a group of related data points that, as a whole, deviate from the normal pattern, even if individual points may not be anomalous [32]. In energy data, collective anomalies can manifest as coordinated attacks or cascading failures, which are particularly challenging to detect.

Recent studies have examined the use of temporal dependencies and spatial correlations to enhance the detection of contextual and collective anomalies. For example, graph neural networks (GNNs) have been utilized to model the spatial relationships between different substations, enabling the detection of collective anomalies that span multiple locations [89], [90]. Furthermore, sequence modelling techniques, including recurrent neural networks (RNNs) and transformers, have been utilized to capture the temporal dependencies in energy consumption data, improving the detection of contextual anomalies that evolve overtime.

2.3.1.2 Supervised, Semi-Supervised, and Unsupervised Approaches

Anomaly detection algorithms can be broadly classified as supervised, semi-supervised, or unsupervised, depending on the availability of labelled data [91]. Supervised methods require labelled training data, including both normal and anomalous samples, which can be challenging to obtain in energy systems due to the rarity of anomalies. Semi-supervised methods leverage labelled normal data and unlabelled data, while unsupervised methods rely solely on unlabelled data. Supervised methods, such as deep neural networks trained with labelled anomalies, can achieve high detection accuracy when sufficient labelled data is available [1].

Given the frequent scarcity of labelled anomaly data, semi-supervised or unsupervised anomaly detection methods are often employed. Semi-supervised techniques, like one-class SVMs and autoencoders, learn a representation of normal data and identify anomalies as deviations. Unsupervised techniques, such as clustering and density-based methods, detect anomalies as outliers that do not conform to the prevalent data patterns. Recent research has focused on developing hybrid approaches that utilize the strengths of different learning paradigms. For instance, semi-supervised deep learning models that leverage both labelled normal data and unlabelled data have shown favourable results in detecting anomalies in energy consumption data [92]. Furthermore, transfer learning techniques have been studied to leverage pre-trained models from related domains, reducing the need for extensive labelled data in the target domain.

2.3.1.3 Evaluation Frameworks for Anomaly Detection

Assessing the efficacy of anomaly detection algorithms is paramount for their practical utility. Nevertheless, the intrinsic class imbalance of energy consumption data and the variety of anomaly patterns complicate performance evaluation. Traditional metrics, including accuracy, can provide a skewed assessment in imbalanced datasets due to the prevalence of normal instances. Therefore, metrics like precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) are more appropriate for evaluating anomaly detection capabilities [93]. Furthermore, the evaluation framework should consider the trade-off between detection accuracy and computational efficiency, as real-time anomaly detection is often required in energy systems.

Recent research has emphasized the need for comprehensive evaluation frameworks that consider various aspects of anomaly detection, including detection latency, false alarm rate, and robustness to noise and adversarial attacks. Furthermore, the use of realistic simulation environments and benchmark datasets is essential for evaluating the performance of anomaly detection algorithms under diverse operating conditions [94] Additionally, the cost of false positives versus false negatives should be considered. In a system, a false negative might mean a power outage, and a false positive might mean wasted time from an operator. Depending on the system, one of these might be more costly than the other.

In summary, anomaly detection in energy consumption data is a critical task that requires a multi-faceted approach. Future research ought to focus on the advancement of robust and adaptable algorithms that can handle the complexity and imbalanced nature of energy data. Furthermore, the development of comprehensive evaluation frameworks and benchmark datasets is essential for advancing the field of anomaly detection in energy systems.

2.3.2 Deep Learning Architectures

Deep learning, a paradigm rooted in the construction of multi-layered neural networks, has profoundly reshaped data analysis across numerous domains [95]. Its capability in learning hierarchical representations from complex datasets makes it particularly promising for anomaly detection in energy consumption data, a domain characterized by intricate temporal and spatial dependencies, and significant class imbalance. The fundamental building blocks of deep learning are neural networks, composed of interconnected nodes that are organized into input, hidden, and output layers[96]. The introduction of non-linearities through activation functions, such as ReLU or sigmoid enables these networks to model complex relationships [95], [97]. Training these networks relies heavily on backpropagation and gradient descent, adjusting weights and biases iteratively to minimize the difference between predicted and actual outputs [98].

2.3.2.1 Convolutional Neural Networks (CNNs) for Spatial Feature Extraction

Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in spatial feature extraction, particularly in image processing [55], [98]. In the context of energy systems, CNNs can be adapted to analyze spatial patterns of energy

consumption, such as those arising from geographically distributed smart meters or substations. By treating spatial energy consumption data as an image-like structure, CNNs can learn hierarchical features that capture local and global spatial dependencies. For instance, CNNs can identify patterns of correlated energy consumption across neighbouring substations, which may indicate collective anomalies or cascading failures.

Recent research has explored the use of 1D and 2D CNNs for time series anomaly detection, transforming time series data into image representations to leverage the spatial feature extraction capabilities of CNNs [99]. This approach can be particularly useful for detecting anomalies that exhibit spatial correlations, such as those arising from distributed attacks or equipment failures spanning multiple locations. However, the effectiveness of CNNs in energy data depends on proper data representation and careful tuning of convolutional filters and pooling layers.

2.3.2.2 Recurrent Neural Networks (RNNs) for Temporal Dependencies

Recurrent Neural Networks (RNNs) are designed to handle sequential data, making them well-suited for capturing temporal dependencies in energy consumption data [55],[98] RNNs maintain hidden states that store information about past inputs, enabling them to model time-varying patterns and detect anomalies that evolve overtime. However, a key limitation of traditional RNNs is the vanishing gradient problem, hindering their ability to model long-range temporal dependencies. Long Short-Term Memory (LSTM) networks were developed to overcome this issue by introducing memory cells and gating mechanisms. These components regulate information flow and gradient propagation, thereby enabling LSTMs to effectively capture and learn long-term dependencies in sequential data [100].

LSTMs have been widely used in time series anomaly detection, demonstrating superior performance compared to traditional RNNs. They can learn complex temporal patterns in energy consumption data, such as seasonal variations and daily cycles, and detect deviations from these patterns that indicate anomalies. Bidirectional LSTM (BiLSTM) networks further enhance the capabilities of LSTMs by processing input sequences in both forward and backward directions [101]. This enables the network to capture contextual information from both preceding and subsequent time steps, thereby, improving the detection of anomalies that depend on both preceding

and succeeding events. In energy consumption data, BiLSTMs can effectively capture complex temporal dependencies, such as those arising from coordinated attacks or demand fluctuations that span multiple time intervals. Recent studies have shown that BiLSTMs outperform unidirectional LSTMs in time series anomaly detection tasks, particularly when dealing with long-range dependencies and complex temporal patterns [95],[102].

2.3.2.3 Transformer Architectures and Attention Mechanisms

Transformer architectures, originally developed for natural language processing, have recently gained popularity in time series anomaly detection [103], [104]. Transformers rely on attention mechanisms, that permits the model to selectively focus on pertinent parts of the input sequence, capturing long-range dependencies without the limitations of RNNs. Attention mechanisms compute attention weights that indicate the importance of different time steps, aiding the model to learn complex temporal relationships. In energy consumption data, transformers can effectively capture long-range dependencies, such as those arising from seasonal variations or long-term trends, and detect anomalies that deviate from these patterns.

Recent research has explored the use of transformer-based models for time series anomaly detection, demonstrating superior performance compared to RNNs and LSTMs, particularly when dealing with long sequences and complex temporal patterns [105], [96],[97]. The self-attention mechanism within transformers allows the model to learn the relationships between different time steps without relying on recurrent connections, which can be computationally expensive.

2.3.2.4 Graph Neural Networks (GNNs) for Interconnected Energy Systems

Graph Neural Networks (GNNs) are designed to process graph-structured data, making them appropriate for modelling interconnected energy systems, such as power grids and smart meter networks [89]. GNNs can capture the spatial relationships between different components of the energy system, such as substations and transmission lines, and detect anomalies that arise from interactions between these components. By modelling the energy system as a graph, GNNs can capture spatial dependencies and detect collective anomalies that span multiple locations. Recent research has explored the use of GNNs for anomaly detection in power grids, demonstrating their ability to identify coordinated attacks and cascading failures.

GNNs can also be used to model the spatial relationships between smart meters, enabling the detection of anomalies that arise from compromised devices or malicious activities [98]. The potential to GNNs to model complex spatial dependencies makes them particularly useful for detecting anomalies in interconnected energy systems.

2.3.2.5 Autoencoders

Autoencoders represent unsupervised learning models designed for feature learning and dimensionality reduction [25]. This method is especially useful in unsupervised learning scenarios. Comprising an encoder and a decoder, autoencoders reconstruct input data, serving applications like data compression and anomaly detection. The application of Variational Recurrent Autoencoders (VRAEs) with attention mechanisms also presents a promising approach to smart meter data anomaly detection [64]. VRAEs, by combining the generative capabilities of variational autoencoders with the sequential processing of recurrent neural networks, effectively capture both the probabilistic distribution and temporal dependencies inherent in energy consumption patterns. Furthermore, the incorporation of attention mechanisms enables the model to selectively focus on the most pertinent time steps, improving its ability to identify subtle anomalies within complex time series data. This methodology proves particularly valuable in addressing challenges posed by noisy, unlabeled, and highly variable smart meter data, enabling the robust detection of deviations from normal consumption behaviour that may indicate faults, cyberattacks, or other operational irregularities.

2.3.2.6 Generative Adversarial Networks

Generative Adversarial Networks (GANs), on the other hand, pit a generator against a discriminator to create new data samples, contributing to tasks such as image generation and style transfer [62], [10]. Generative Adversarial Networks can learn the fundamental distribution of normal energy consumption data and identifying deviations as anomalies. GANs operate in an adversarial manner, where the generator attempts to create realistic energy consumption patterns, while the discriminator distinguishes between generated and real data [106]. In the context of anomaly detection, the discriminator, trained on normal data, are able to recognize anomalous patterns that exhibit a large deviation from the learned distribution. The generator, conversely, can be used to synthesize realistic, but rare, anomalous data to augment

training sets, addressing the challenge of imbalanced datasets [106]. This capability makes GANs particularly valuable in scenarios where labeled anomaly data is scarce, or where the aim is to detect novel, previously unseen anomalies. However, training GANs can be computationally demanding and necessitates meticulous hyperparameter tuning to mitigate issues such as mode collapse, which can impact the accuracy of anomaly detection.

2.3.2.7 Comparative Strengths and Limitations of Different Architectures

Each deep learning architecture possesses distinct strengths and limitations, rendering it variably suitable for different anomaly detection tasks. CNNs excel in spatial feature extraction but may be suboptimal for capturing long-range temporal dependencies. RNNs and LSTM networks are well-suited for modeling temporal dependencies, though they can face challenges with extended sequences and intricate temporal patterns. BiLSTMs improve the capabilities of LSTMs by processing input sequences in both directions, improving the detection of anomalies that depend on both past and future events. Transformer architectures offer superior performance in capturing extended dependencies and intricate temporal patterns, but may be computationally expensive. GNNs are well-suited for modelling interconnected energy systems, but may require careful design of the graph structure and node features.

In the context of imbalanced energy consumption data, the choice of architecture should consider the specific characteristics of the data and the type of anomalies to be detected. For instance, if the anomalies exhibit spatial correlations, CNNs or GNNs may be more appropriate. If the anomalies exhibit complex temporal patterns, RNNs, LSTMs, BiLSTMs, or transformers may be more suitable. Furthermore, ensemble methods that combine the strengths of different architectures can be used to improve the overall detection performance. For example, an ensemble of CNNs and LSTMs can capture both spatial and temporal dependencies, improving the detection of anomalies that exhibit both spatial and temporal characteristics.

Hyperparameter tuning becomes an essential aspect, influencing factors like learning rates, network architecture, and regularization techniques[24]. Deep learning frameworks like TensorFlow, PyTorch, and Keras make it easier to build and train these complex models [65]. These frameworks provide tools and abstractions, streamlining the development process for practitioners and researchers [65]. The deep

learning model’s interpretability poses a challenge due to their complexity. The selection of an appropriate deep learning architecture for anomaly detection in imbalanced energy consumption data requires careful consideration of the spatial and temporal characteristics of the data, the type of anomalies to be detected, and the computational resources available. Future research should prioritize developing hybrid architectures and ensemble methods that integrate the strengths of diverse deep learning models to enhance the accuracy and robustness of anomaly detection systems.

2.3.3 Explainability and Interpretability Methods

Robust explainability and interpretability techniques must be developed and implemented due to the growing complexity of deep learning models, especially deep ensemble models used for anomaly detection in imbalanced energy consumption data. These techniques are essential for comprehending how these models make decisions, establishing confidence in their predictions, and making it easier to find useful insights. Explainability is not just a desirable feature but a basic necessity in the context of critical infrastructure, such as energy systems, where missed anomalies or false alarms can have serious repercussions.

2.3.3.1 Local Interpretability Methods (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is a prevalent local interpretability technique that approximates any machine learning model with a locally faithful, interpretable model to explain its predictions. [107]. LIME explains individual predictions through input perturbations around a specific instance and observation of changes in the model's output. It then trains a simple, interpretable model (e.g., a linear model) on this perturbed data, using the original model's predictions as labels. This local model elucidates the original model's behaviour for that particular instance. In the context of energy consumption anomalies, LIME can identify the specific features or time points driving the model's anomaly prediction. For example, LIME can highlight the sudden increase in energy consumption at a particular time point that led the model to classify it as an anomaly. LIME's model agnostic nature makes it very useful for ensemble models, where individual models may be very different in their internal structure.

2.3.3.2 Global Interpretability Approaches (SHAP)

SHapley Additive exPlanations (SHAP) is a global interpretability technique that aims

to explain the overall machine learning model's behaviour by assigning importance values to each feature [108]. SHAP values, rooted in game theory, quantify each feature's contribution to a model's prediction by considering all possible feature combinations. SHAP values offer a consistent and theoretically grounded measure of feature importance, rendering them particularly valuable for elucidating the global behaviour of complex models [109]. In energy anomaly detection, SHAP can be used to identify the most important features that contribute to the model's predictions of anomalies. For example, SHAP can reveal that certain time-of-day features or weather conditions are consistently associated with anomalies. SHAP's global approach is highly beneficial for understanding the overall behaviour of ensemble models by identifying the most salient features across all constituent models.

2.3.3.3 Feature Importance Visualization Techniques:

Feature importance visualization techniques visually represent the significance of different features in a machine learning model, aiding in the identification of influential features and providing insights into the model's decision-making process [110]. Common feature importance visualization techniques include bar charts, heatmaps, and feature importance plots. In the context of energy consumption anomalies, feature importance visualization can help identify the features that are most indicative of anomalies, such as specific time points, sensor readings, or weather conditions. These visualizations can also be used to compare the feature importance across different models in an ensemble, providing insights into the diversity and complementarity of the ensemble members. Partial dependence plots (PDPs) and individual conditional expectation (ICE) plots can also visualize the relationship between individual features and model predictions [111].

2.3.3.4. Attention-based Interpretability in Deep Learning Models

Attention mechanisms, which are increasingly used in deep learning models for time series analysis, provide a natural way to interpret the model's focus on different parts of the input sequence [112]. Attention weights can be visualized to identify the time points or features that the model deems most relevant for its predictions. In energy anomaly detection, attention-based interpretability can help identify the specific time points or sensor readings that contributed to the model's prediction of an anomaly [113] For example, attention weights can highlight the sudden increase in energy

consumption at a particular time point that led the model to classify it as an anomaly. Attention visualizations can also reveal the temporal dependencies that the model has learned, providing insights into the dynamics of energy consumption patterns. Furthermore, the attention weights can be used to identify the most influential models within an ensemble, highlighting the models that are most relevant for a particular prediction.

2.3.3.5 Model-agnostic vs. Model-specific Interpretability Techniques

Interpretability techniques can be broadly classified as model-agnostic or model-specific. Model-agnostic techniques, such as LIME and SHAP, can be practical to any machine learning model, irrespective of its internal structure. Model-specific techniques, on the other hand, are designed for specific types of models, such as attention-based interpretability for deep learning models with attention mechanisms. Model-agnostic techniques are particularly useful for interpreting complex models, such as deep ensemble models, where the internal structure may be difficult to understand [114]. Model-specific techniques offer detailed insights into a model's behavior but are limited to particular model types. In energy anomaly detection, both model-agnostic and model-specific techniques can elucidate the decision-making process of deep ensemble models. For example, LIME and SHAP can be used to identify the most important features and time points, while attention-based interpretability can provide insights into the temporal dependencies learned by the model [115].

In the context of imbalanced energy data, it is imperative to take into account the impact of class imbalance on interpretability. For instance, feature importance visualizations may be biased towards the majority class, highlighting features that are indicative of normal consumption patterns but not necessarily of anomalies. Consequently, it is essential to use interpretability techniques which are robust to class imbalance, for instance those that consider the contribution of features to the prediction of both normal and anomalous samples. Additionally, it is important to evaluate the effectiveness of interpretability methods using appropriate metrics, such as the accuracy of feature importance rankings and the faithfulness of local explanations. In the case of ensemble models, methods that can explain the overall behaviour of the ensemble as well as the individual models are highly valuable.

To sum up, explainability and interpretability methods are indispensable for building trust in deep ensemble models for anomaly detection in imbalanced energy consumption data. The use of local and global interpretability techniques, feature importance visualization, and attention-based interpretability can contribute to a deeper understanding of the model's decision-making logic, facilitating the identification of actionable insights and improving the reliability of energy systems.

2.3.4 Ensemble Learning Approaches

In machine learning, ensemble methods represent a powerful approach to improving predictive model performance by integrating the strengths of multiple individual models [30]. The fundamental concept underlying ensembles is the collaboration of diverse models, each contributing its unique perspective to create a more robust and accurate collective prediction [29]. These ensemble methods come in various forms, with bagging, boosting, and stacking being the prominent types. Bagging entails training models independently on distinct subsets of data, generally via bootstrap sampling, and then aggregating their predictions using techniques like averaging or voting [116].

Random Forest, a well-known ensemble model, exemplifies this approach by constructing a multitude of decision trees[117]. Boosting, on the other hand, takes an iterative approach, where weak models are trained successively, with each subsequent model focused on fixing errors made by its predecessors. Algorithms like AdaBoost and Gradient Boosting Machines showcase the boosting paradigm, effectively adjusting the model's emphasis on misclassified instances during each iteration[116]. Stacking introduces the notion of a meta-learner, a higher-level model that integrates predictions from several base models.

This meta-learner is trained on the outputs of the base models, essentially learning how to best weigh their contributions. Stacking, with its emphasis on leveraging the complementary strengths of diverse models, has found success in various machine learning applications [66]. The efficiency of ensemble methods relies heavily on diversity among base models. This diversity can be accomplished by training models on diverse subsets of the data., using distinct features, or employing various architectures [30]. The goal is to ensure that individual models capture different aspects of the underlying patterns in the data, collectively enhancing the ensemble's

ability to generalize well [30],[66]. The combination of these diverse models often leads to improved performance, robustness, and better generalization to unseen data. Ensemble methods are versatile and applicable across different types of machine learning tasks, including classification, regression, and even unsupervised learning scenarios [30],[66]. Hyperparameter tuning is a crucial aspect of working with ensemble methods, determining factors such as the number of base models, learning rates in boosting, or meta- learner specifications in stacking [30],[66].

A trade-off exists between ensemble size and computational costs, as larger ensembles may offer performance benefits up to a certain point but come with increased resource requirements [30]. Out-of-Bag (OOB) error estimation is a unique feature in bagging methods like Random Forest. It offers a crucial evaluation of a model's generalization performance without requiring a separate validation set, aiding in the evaluation and fine-tuning of ensemble models [66]. Ensemble methods, despite their undeniable advantages, may introduce challenges in interpretability, especially as the complexity of the ensemble increases [30]. Optimizing the trade-off between performance and interpretability is an ongoing consideration in the application of ensemble methods, with techniques like feature importance analysis providing insights into the ensemble's decision-making process [66].

Ensemble modeling approaches, particularly those focused on fusion strategies, model selection, and hyperparameter optimization, are pivotal for enhancing anomaly detection in imbalanced energy consumption data. Fusion strategies, at their core, dictate how the predictions of individual models within the ensemble are combined to produce a final prediction. Voting, a straightforward strategy, aggregates predictions based on majority rule, suitable for classification tasks where a clear consensus is expected [118]. Averaging, on the other hand, computes the mean or weighted mean of predictions, often preferred for regression tasks or when dealing with continuous anomaly scores. Learning-based fusion strategies, such as stacking or blending, train a meta-learner to combine the outputs of base models, allowing for more complex and adaptive fusion patterns [119]. In the context of energy data, where anomalies can manifest in subtle and varied ways, a learning-based fusion strategy might be particularly advantageous, as it can tailor to the specific properties of the anomalies and the strengths of the individual models.

Model selection within an ensemble framework is critical for ensuring diversity and maximizing performance. Selecting models with complementary strengths can improve the ensemble's potential to capture multiple facets of the data and reduce the risk of overfitting. For example, in energy consumption anomaly detection, an ensemble might include models that excel at capturing temporal dependencies (e.g., LSTMs) and models that excel at capturing spatial relationships (e.g., CNNs). Methods like cross-validation and out-of-bag error estimation can be leveraged to evaluate the performance of individual models and select the most promising candidates for the ensemble [120]. Furthermore, diversity-promoting techniques, such as bagging and boosting, can be used to generate diverse sets of models, further enhancing the ensemble's performance. In imbalanced datasets, it is important to select models that are robust to class imbalance, such as those that employ cost-sensitive learning or resampling techniques.

Hyperparameter optimization is paramount for achieving maximal performance in both constituent models and the integrated ensemble. Methodologies such as grid search (an exhaustive parameter space traversal), random search (stochastic sampling from defined parameter distributions), and Bayesian optimization (a probabilistic search strategy guided by performance likelihood) are employed to identify optimal hyperparameter configurations [121],[122]. In ensemble models, it is important to optimize not only the hyperparameters of individual models but also the hyperparameters of the fusion strategy [123] For instance, in a learning-based fusion strategy, the hyperparameters of the meta-learner must be optimized to ensure effective combination of the base model predictions. Additionally, methods that optimize hyperparameters for the ensemble as a whole, rather than individually, can lead to better overall performance. Methods like evolutionary algorithms have been harnessed for this purpose [124]. In the context of imbalanced energy data, hyperparameter optimization should consider the impact of class imbalance on model performance, ensuring that the models are optimized for minority class detection.

The choice of fusion strategy, model selection method, and hyperparameter optimization technique should be guided by the specific characteristics of the energy consumption data and the nature of the anomalies to be detected. For instance, if the anomalies are highly variable and exhibit complex temporal patterns, a learning-based

fusion strategy and diverse model selection approach may be more appropriate. If the anomalies are rare and the data is highly imbalanced, hyperparameter optimization techniques that are robust to class imbalance should be used. Furthermore, the computational resources available should also be considered when selecting these techniques, as some methods, such as Bayesian optimization and evolutionary algorithms, can be computationally expensive.

In essence, ensemble methods exemplify the collective strength derived from the amalgamation of diverse models, demonstrating the synergy achieved when multiple perspectives contribute to a unified and enhanced predictive model. As the field of machine learning evolves, ensemble methods remain at the forefront, offering a robust and versatile toolkit for addressing a wide array of complex predictive modeling challenges [30],[66].

2.3.5 Class Imbalance Handling Techniques

The challenge of class imbalance, where normal energy consumption data significantly outweighs anomalous data, poses a critical obstacle to developing effective anomaly detection systems. This section critically evaluates the literature on class imbalance handling techniques, focusing on data-level, algorithm-level, and hybrid approaches, and examining the appropriateness of performance metrics for imbalanced datasets.

2.3.5.1 Data-Level Approaches

Data-level approaches, primarily sampling techniques, aim to balance the class distribution by modifying the training dataset. Oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), generate synthetic samples for the minority class, increasing its representation [27], While effective in mitigating imbalance, SMOTE can introduce synthetic noise and overgeneralization, particularly in high-dimensional spaces. Undersampling techniques, like random undersampling, reduce the majority class samples but can lead to information loss and bias [125]. Recent advancements have explored hybrid sampling methods, combining oversampling and undersampling, such as SMOTE-Tomek and SMOTE-ENN, to address the limitations of individual techniques. However, the computational cost of these hybrid methods can be significant, especially for large-scale energy datasets. Moreover, sampling techniques, especially oversampling, can distort the original data

distribution, potentially leading to models that do not generalize well to unseen data. In scenarios where the minority class is very rare, oversampling may not be enough to provide the model with sufficient information to learn the anomalies.

2.3.5.2 Algorithm-Level Approaches

Algorithm-level approaches modify the learning algorithm to account for class imbalance. Cost-sensitive learning, a key technique, assigns differential misclassification costs to classes, imposing a higher penalty for errors in the minority class[126]. This approach can be implemented by modifying the loss function or adjusting the decision threshold of the classifier. Cost-sensitive learning can be effective in improving minority class detection, but requires careful selection of cost parameters, which can be challenging and data-dependent. Recent research has explored adaptive cost-sensitive learning methods that dynamically adjust cost parameters based on the characteristics of the data and the learning process [127]. However, the effectiveness of cost-sensitive learning depends on the potential of the algorithm to effectively incorporate the cost information, which may not be straightforward for all deep learning architectures. Furthermore, cost-sensitive learning assumes that the cost of misclassification is known or can be accurately estimated, which may not always be the case in real-world energy systems [128].

2.3.5.3 Hybrid Approaches

Hybrid approaches combine data-level and algorithm-level techniques to leverage the strengths of both. For example, cost-sensitive learning can be applied to a dataset that has been pre-processed using oversampling or undersampling [129]. Hybrid approaches can achieve superior performance compared to individual techniques but require careful selection and integration of the component methods. Recent research has focused on developing adaptive hybrid approaches that dynamically select and combine different techniques based on the characteristics of the data and the learning process [130]. However, the complexity of these hybrid methods can make them difficult to implement and interpret. Furthermore, the computational overhead of applying combined techniques can be significant.

2.3.5.4 Performance Metrics for Imbalanced Data

Traditional metrics, such as accuracy, can yield misleading evaluations in imbalanced datasets due to potential majority class dominance. Precision (positive predictive

value), recall (sensitivity), F1-score (the harmonic mean of precision and recall), and the area under the receiver operating characteristic curve (AUC-ROC) are more rigorous metrics for evaluating minority class detection performance [131]. Specifically, precision quantifies the proportion of true positives among all positive predictions, while recall quantifies the proportion of true positives among all actual positives. The F1-score provides a balanced assessment of these two measures [132]. AUC-ROC assesses the model's discriminatory capacity between normal and anomalous instances across a spectrum of decision thresholds.

Recent research has investigated specialized metrics like the geometric mean (G-mean) and balanced accuracy to address limitations of traditional metrics [133]. However, metric selection should align with the specific objectives of the anomaly detection task. For energy data, where false negatives can have critical consequences, metrics emphasizing recall may be more pertinent. Furthermore, the costs associated with false positives and false negatives warrant consideration during model evaluation. Finally, the severity of class imbalance must be factored into the evaluation, as high AUC-ROC values may be insufficient in extremely imbalanced scenarios.

In conclusion, addressing class imbalance is crucial for developing effective anomaly detection systems for energy consumption data. Data-level, algorithm-level, and hybrid approaches offer various strategies for mitigating the imbalance effect, but each has its own strengths and limitations. The selection of appropriate techniques and performance metrics should be predicated on the specific characteristics of the data and the objectives of the anomaly detection task. Future research should focus on developing adaptive and robust techniques that can effectively handle the challenges posed by imbalanced energy data.

2.4 Empirical Literature

This study utilised empirical evidence gathered through experiments to address its research questions.

2.4.1 Energy Consumption Anomaly Detection Studies

This section examines the existing literature that focuses on detecting deviations from typical patterns of energy use. To create a thorough grasp of the state-of-the-art, this section analyses various approaches, datasets, and assessment metrics used in previous research. A concise overview of the diverse models utilised to tackle class

imbalance in energy consumption anomaly detection, detailing their specific datasets, training methodologies, and corresponding performance metrics.is described. This analysis lay the groundwork for the development of a strong deep ensemble anomaly detection model.

‘A combined genetic optimization with AdaBoost ensemble model for anomaly detection in buildings electricity consumption “ [134]

This study proposes a genetic algorithm-optimized ensemble model for electricity theft detection. Initially, the SMOTE is employed as an auxiliary training approach to generate synthetic samples that approximate the distribution of real data at a general level.

This additive and deep history characteristic underpins the modern efficacy of the AdaBoost ensemble learning algorithm, offering reduced computational cost, enhanced predictive accuracy, and improved robustness. Subsequently, Principal Component Analysis (PCA) is applied during preprocessing to extract anomalous components of electricity consumption through dimensionality reduction analysis. Finally, an AdaBoost-based ensemble deep learning network is developed to capture inherent information from extended time series data and identify abnormal consumption patterns among labelled users. Furthermore, the hyperparameters of this deep neural network are adjusted using a genetic algorithm (GA).

The experimental findings prove that the proposed model exhibits higher sensitivity and Area Under the Curve (AUC) compared to other methods, including Support Vector Machines (SVM), Random Forests (RF), and traditional Artificial Neural Networks (ANN), when evaluated on electricity consumption data from smart meters belonging to the SGCC[134]. However, the method's performance remains partially dependent on a relatively limited number of users with subtle distinguishing characteristics. For instance, instances involving minor electricity theft can be effectively concealed, and normal users exhibiting frequent load fluctuations (Tload) due to events like business trips may be misclassified as electricity thieves. To address these limitations, future research could focus on refining hyperparameter optimization techniques, training methodologies, and feature extraction methods.

“An adaptive synthesis to handle imbalanced big data with deep Siamese network for electricity theft detection in smart grids “ [135].

This study addresses the class imbalance problem by employing Adaptive Synthetic Sampling (ADASYN). To discern the characteristics of legitimate and fraudulent energy consumers, deep Siamese networks (DSNs) with a composite architecture of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) are utilized. The CNN module is designed to extract spatial features from weekly energy consumption snapshots, while the LSTM module facilitates sequential pattern learning. Subsequently, the DSN performs classification based on shared representations derived from both the CNN and LSTM modules. Real-time smart meter data is analyzed using varying train-test split ratios. Simulation results indicate the proposed model's superior performance in terms of Area Under the Curve (AUC), F1-Score, precision, and recall.

However, Deep Siamese Networks, by focusing on learning localized feature relationships within example pairs, may encounter limitations in capturing global patterns across multiple features inherent in energy consumption data. Anomalous behaviours can manifest through novel and intricate interactions that may not be readily discernible by the pairwise comparison approach of Siamese Networks, which also emphasizes learning from proximal instances. Furthermore, the CNN architecture, originally conceived for spatial data analysis, may not inherently represent temporal variations present in time series data, a common attribute in energy consumption analysis. Consequently, temporal anomaly patterns might not be entirely represented by a standard CNN.

While LSTMs excel at capturing sequential dependencies, they may lack the capacity to encapsulate macroscopic global patterns across all features within energy consumption data. As a result, subtle anomalous behaviours that do not manifest as strong sequential deviations might be missed by LSTMs.

“Anomaly Prediction in Electricity Consumption Using a Combination of Machine Learning Techniques “[92]

This research investigates a two-stage anomaly detection framework for smart meter

electricity consumption data. Initially, an Isolation Forest algorithm is applied to classify individual power usage readings as either normal or abnormal, producing a sequence of labelled data points.

Subsequently, Random Forest and Decision Tree algorithms are employed as predictive models to forecast the occurrence of power anomalies based on these generated label sequences. Empirical evaluation of this framework reveals a consistent ability to detect and predict abnormal consumption events with a 30-minute lead time. Importantly, the Random Forest performance and Decision Tree algorithms exhibited no statistically significant difference across various smart meter datasets, training set sizes, and the length of the input data sequences.

The inclusion of this intermediate classification layer allows for the automated annotation of data instances as normal or anomalous and subsequently enables prediction based on the resulting label sequence. This is achieved without directly integrating the inherent dynamic properties of power consumption data into the predictive modelling stage. However, a recognized limitation of Random Forest algorithms is their potential for bias towards the majority class in scenarios involving imbalanced datasets. The learning process can be disproportionately influenced by the dominant normal class, which may result in a model optimized for normal instances but less proficient in identifying anomalies.

Furthermore, the inherent training mechanisms of Random Forests may lead to a reduced sensitivity towards the minority class, representing the anomalous instances. As an algorithm aiming to maximize overall classification accuracy, Random Forests may struggle to effectively learn the characteristic patterns of infrequent anomalies. Beyond computational efficiency, the presence of rare and intricate anomaly patterns can pose a significant challenge to the learning capabilities of Random Forests, potentially leading to an ensemble that insufficiently captures the subtle distinctions of these infrequent events.

“Electrical Network Fraud Detection Using a CNN and LSTM Model “ [136]

This model is a Deep Learning model composed of a Convolutional Neural Network Distributed in Time (CNN) as well as Long Short-Term Memory (LSTM). Its

objective is to explore existing fraud detection techniques, examine the practicability and effects of CNNT+LSTM models in fraud detection.

It's a time-distributed convolutional layer encapsulated in a 64-layer 1D convnet configuration, followed by max-pooling operation, producing flat features to feed into the LSTM Layer with 50 rows (Units). The output layer was one with a sigmoid activation function. This way the model learns to assign probabilities for both kinds of results. Compilation of the model used a 0.01 learning rate, the Adam optimizer, and binary cross-entropy loss.

The data was shaped so as to make the training subset 80% and validation subset 20% respectively. It was initially considered to use Dropout as a form of regularization, but it was noted that the model's performance was actually better without it.

Data equalization was achieved by oversampling the irregular instances. But in datasets with a small number of anomaly instances, CNNs can end up spending a lot of time on the data that are already present. As a result, CNNs in imbalanced datasets may impede effective learning of the minority class for them to effectively learn particular features of anomalies. The CNN model can still attain high accuracy through prediction of the majority class while paying less attention to the role of class (anomalies), resulting in reduced effectiveness for anomaly detection.

Repeated duplication of minority class instances during oversampling can lead to overfitting. Instead of learning broad-based patterns that can be applied across all data, the model will internalize each case of oversampled examples as if it were memorizing them and so perform poorly on new samples not previously seen.

Moreover, oversampling means more representation of minority class instances, but maybe less diversity within the minority class. If anomalies manifest themselves in various ways, oversampling methods relevant to specific data points might fail to capture all possible varieties.

“Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data” [137]

The electricity consumption data in this model undergoes preprocessing via

interpolation, three-sigma rule-based outlier removal, and normalization. Class imbalance is addressed using the ADASYN algorithm, which oversamples the minority class. The resulting balanced dataset is then fed into a Visual Geometry Group (VGG-16) module for detecting anomalous electricity usage. Classification is performed by an Extreme Gradient Boosting (XGBoost) model optimized with the Firefly Algorithm (FA-XGBoost). Performance evaluation is conducted through simulations and comparative analysis against Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Logistic Regression (LR) models, employing the F1 score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic Area Under the Curve (ROC-AUC), Precision-Recall Area Under the Curve (PR-AUC), and precision as evaluation metrics.

First, simulation results show that when using the projected ADASYN method with the FA-XGboost classifier, the classifier scored an F1 - score of 93.7%, precision of 92.6%, and recall rate up to 97%. Second, the VGG-16 module trained and tested accuracy was 87.2% and 83.5%, which has greater overall performance.

Thirdly, the proposed FA-XGBoost correctly identifies who the true electricity thieves are, that is, testimony to this predictive power success story! in the recall rate reaches 97%. However, based models like XGBoost could have a bias towards the majority class in an imbalanced dataset. A potential limitation of the algorithm is its inherent prioritization of learning patterns prevalent in the majority class, often prioritizing the maximization of majority class accuracy representing anomalies. Additionally, XGBoost's default configuration may exhibit reduced sensitivity towards the minority class, hindering its ability to effectively discern the subtle variations within rare anomaly patterns. Consequently, the strategic assignment of weights to distinct classes during the XGBoost model training phase is likely to yield improved anomaly detection outcomes.

“Electricity Theft Detection in Smart Grids Based on Deep Neural Network “ [138].

Class imbalance is addressed through data interpolation and synthetic data generation. This study analyzes and compares temporal and spectral domain features, employing Principal Component Analysis (PCA) for feature extraction. Bayesian optimization and adaptive moment estimation are utilized for hyperparameter optimization.

Validation results demonstrate a 97% Area Under the Curve (AUC), a 1% improvement over existing works, and 91.8% accuracy, the second-highest benchmark achieved.

However, PCA's focus on maximizing data variance might not align with the discriminative features of anomalies, which can be overshadowed by dominant normal patterns in imbalanced datasets, thus limiting PCA's effectiveness for anomaly detection. Furthermore, deep neural networks are prone to overfitting, particularly with imbalanced classes, where the model might learn noise or specific training patterns, hindering generalization to new data.

“Ensemble machine learning models for the detection of energy theft “ [8]

This research explores ensemble machine learning (ML) for energy theft detection in smart grids via customer consumption patterns. Implemented and evaluated were adaptive boosting, category boosting, extreme gradient boosting, light gradient boosting, random forest, and extra trees to improve prediction by reducing variance and bias. To mitigate overfitting, minority oversampling was applied. The experimental results showed that bagging-based ensemble models outpaced other evaluated algorithms. Specifically, random forest and extra trees achieved the maximum Area Under the Curve (AUC) score of 0.90. Furthermore, precision analysis indicated that the proposed bagging methods were superior.

However, the study pinpoints the importance of effective hyperparameter tuning for determining optimal ensemble weights, which is a complex but critical step towards maximizing anomaly detection performance and classification accuracy. The research also notes that ensemble models can be vulnerable to noisy or irrelevant features, especially if anomalies introduce such noise. Finally, the effectiveness of the ensemble is contingent upon the diversity of its base models, as limited diversity can hinder its ability to capture a comprehensive range of anomaly patterns and effectively deal with class imbalance.

“Fault detection based on one-class deep learning for manufacturing applications limited to an imbalanced database” [139]

A novel fault detection model is proposed, integrating residual computation, one-class classification (Isolation Forest and OCSVM), and deep learning-based time-series prediction. Four networks—MLP, ResNet, LSTM, and ResNet-LSTM—were trained solely on successful production instances for this prediction task.

The residuals generated from these predictions were then used as refined features for the one-class classification component. Evaluation on actual die-casting mass production data showed substantial gains: overall accuracy increased to 96.0% (from 90.0%), manufacturing defect detection accuracy rose to 96.0% (from 84.0%), and the Area Under the Receiver Operating Characteristics (AUC-ROC) improved to 98.96% (from 87.56%). Notably, while ResNet–LSTM was better at predicting successful production, ResNet performed best in detecting production failures. This highlights the potential of one-class deep learning for extracting informative features from time-series data for one-class defect detection. Nevertheless, the inclusion of OCSVM introduces several limitations. Optimizing its hyperparameters (kernel parameters and the 'nu' parameter) can be challenging and significantly impacts performance, especially in the presence of class imbalance.

Moreover, OCSVM's assumption of a unimodal distribution of normal data can be problematic if the actual distribution is multimodal, potentially hindering its ability to model normal behavior and detect anomalies effectively. Lastly, the effectiveness of OCSVM is heavily reliant on the quality of the engineered features; poorly representative features can lead to suboptimal anomaly detection results.

“Imbalanced learning algorithm based intelligent abnormal electricity consumption detection ’’[13]

This paper explores a methodology for Automatic Event Categorization (AEC) detection that initially examines two efficient algorithms through the lens of data balancing (using K-means clustering and SMOTE) and data weighting, employing a kernel extreme learning machine (KELM)-trained artificial neural network (ANN).

The study further investigated more deep representation network-based ELM (EH-DrELM), which is an enhanced multiclass AdaBoost imbalanced learning algorithm (AdaBoost-ID), and the deep weighted ELM (DWELM). A critical consideration

regarding this approach is the use of SMOTE for data balancing. SMOTE's inherent assumption of feature independence poses a challenge, as it can generate synthetic instances that do not accurately reflect the minority class distribution when features exhibit correlations.

Furthermore, SMOTE may have challenges in effectively capturing the nuanced and infrequent patterns associated with anomalies, particularly if these anomalies are infrequent or have diverse characteristics, potentially leading to synthetic data that underrepresents their true complexity. Lastly, SMOTE's vulnerability to noisy data within the minority class is a concern, as it can propagate this noise into the generated synthetic samples, potentially impairing the overall performance of the detection algorithms.

“Intelligent Anomaly Detection Method of Gateway Electrical Energy Metering Devices using Deep Learning” [140]

A hybrid deep-learning model combining Stacked Autoencoders (SAE) and LSTM is proposed for intelligent anomaly detection in gateway electrical energy metering devices. This SAE-LSTM model initially utilizes SAE to draw out deep latent features from three-phase voltage data. Subsequently, LSTM leverages these features to distinguish anomalous events. A key strength of this model is its potential to effectively capture the temporal dependencies within the electrical data, yielding improved anomaly detection accuracy. Simulation results under varying signal-to-noise ratios confirm the benefits of the SAE-LSTM approach.

Real-world dataset evaluations further demonstrate its applicability for anomaly detection in practical gateway energy metering systems. Nevertheless, the reliance on SAEs for feature extraction introduces certain limitations.

Highly variable or patternless energy consumption anomalies may hinder SAEs from learning a clear latent representation. Additionally, the static nature of the SAE representation can be a drawback when dealing with dynamic or evolving anomaly patterns, potentially leading to missed novel anomalies.

Furthermore, the assumption of a unimodal distribution in the SAE's latent space may

limit its capacity to effectively represent the diversity of both normal and anomalous data if their distributions are multimodal or complex.

“Real-World Anomaly Detection by Using Digital Twin Systems and Weakly Supervised Learning “[141]

This paper introduces a Digital Twin-augmented training dataset, simulating normal machinery operation with limited real-world labelled anomalies. A Siamese Autoencoder (SAE) architecture, coupled with a cluster centers (CC) strategy, is proposed for anomaly detection in poorly supervised settings.

Evaluated on a real-world facility monitoring dataset against state-of-the-art algorithms using diverse metrics, the proposed SAE-based solutions significantly outperformed benchmarks across hyperparameter settings.

However, SAEs require sufficient anomaly examples for robust similarity metric learning, potentially limiting generalization with scarce anomalies. Extreme class imbalance may bias SAEs towards dominant normal patterns, hindering minority anomaly representation. Furthermore, SAEs may struggle to capture sparse, scattered anomalies due to limitations in the learned similarity metric.

“Self-Adaption AAE-GAN for Aluminium Electrolytic Cell Anomaly Detection” [142]

A novel anomaly detection model utilizes Generative Adversarial Networks (GANs) to model complex high-dimensional image distributions. The self-adaptive AAE-GAN converts multi-dimensional time series data into a 2D matrix and trains solely on normal data. Anomaly detection relies on reconstruction error. Experiments show high accuracy and speed. However, GANs are susceptible to mode collapse, potentially underrepresenting diverse anomalies, particularly in energy data. Controlling realistic generation of rare anomalies is challenging. Class imbalance, with minority anomalies, can bias generation towards normal patterns.

2.4.2 Empirical Comparison of Deep Learning Architectures

The extant literature demonstrates the predominant utilization of Random Forests and Support Vector Machines among classical machine learning methodologies for

tackling class imbalance in the context of energy consumption anomaly detection. Conversely, recent trends indicate a significant increase in the utilization of deep learning models, notably CNNs and LSTM networks. Among all the models, the most widely used training technique is the oversampling minority class. Finally, five metrics are commonly used to assess the numerous deep learning and machine learning models. These consist of precision, recall, accuracy, area under the curve (AUC), and F1 score.

Although the conventional machine-learning-based models have demonstrated success in dealing with class imbalance challenge in the anomaly detection in energy consumption data, the deep-learning-based models have, on the other hand, shown success to a greater extent. Support Vector Machines face the main challenge of struggling to effectively handle sparse anomalies, particularly when these anomalies are scattered across the feature space. The model might fail to capture the nuanced patterns of rare anomalies owing to the focus on maximizing the margin between classes. Random Forests also struggle to effectively detect rare anomalies, especially when the anomalies are sporadic or occur infrequently. The aggregated decision of multiple trees in ensemble methods may inadequately capture the subtle characteristics of rare anomalies. The higher performance of deep learning models over classical approaches in addressing class imbalance is attributed to their capacity for automated learning of intricate representations. However, Convolutional Neural Networks (CNNs) face a primary limitation in effectively capturing sparse anomalies distributed across the feature space.

The hierarchical feature learning process may not focus adequately on learning representations specific to rare anomalies. They also struggle to learn from multimodal distributions of anomalies effectively. If anomalies exhibit complex structures or multiple patterns, CNNs might not capture the full diversity of the data. Long Short-Term Memories (LSTMs) face the main challenge of struggling to effectively learn from multimodal temporal distributions of anomalies. Complex or diverse anomaly patterns across different time scales might pose challenges for LSTM-based anomaly detection. They are also sensitive to class imbalance, especially when anomalies are significantly underrepresented. The model prioritizes learning from the majority class, reducing its sensitivity to anomalies.

Therefore, future research should prioritize the development and rigorous tuning of deep neural networks to effectively capture complex patterns in imbalanced energy consumption data. This approach offers a promising avenue for creating more robust anomaly detection models capable of addressing class imbalance challenges. Moreover, adoption of an effective transfer learning strategy in the training process of these models is also promising. The process of applying knowledge from one task to another is known as transfer learning. It allows the model to transfer learned features from a source task, where anomalies may be more evenly distributed, to a target task. Transfer learning would further leverage the information from well-balanced datasets that would further improve anomaly detection in imbalanced scenarios. This further improves on the deep learning technique through an improved training process so as to create a more effective model that effectively handles the class imbalance challenges in anomaly detection in energy consumption. Moreover, adoption of the ensemble method in the deep models would solve the overfitting challenge where each base model is trained independently, and the diversity helps prevent overfitting because different models may focus on different aspects of the data.

The critical analysis of popular evaluation metrics for class-imbalanced anomaly detection in energy consumption indicates that the optimal metric selection depends on dataset characteristics and the prioritized aspects of model performance. For example, the Precision recall curve has shown greater success since it better reflects the correct prediction of the minority class compared to the other metrics. If false positives or false negatives have different implications, precision or recall might be more critical. Similarly, AUC-ROC or AUC-PR can be effective when the balance between true positives and false positives is crucial. Metrics that take into account unbalanced classes, such as recall, precision, F1 score, AUC-ROC, or AUC-PR, are very important in the detection of anomalies in energy consumption, which are frequently uncommon events. However, it is important to consider selecting multiple metrics to comprehensively evaluate model performance, as no single metric can fully capture the performance in all aspects.

Notably, the ongoing development of techniques to mitigate class imbalance challenges is evidenced by the study of transfer learning, ensemble methods, and specialized architectures like convolutional and recurrent neural networks. As the rich

tapestry of existing research is navigated, it becomes evident that the quest for effective anomaly detection in energy consumption goes beyond. The synthesis of varied perspectives offered in this review sets the stage for the development of novel frameworks that can robustly handle the complexities of class imbalance, ultimately contributing to the advancement of anomaly detection in the dynamic realm of energy consumption systems. Moving forward, this literature review serves as a core tenet for the subsequent phases of this research study. The identified gaps, challenges, and successes pave the way for the formulation and evaluation of innovative anomaly detection models tailor-made specifically for the exceptional characteristics of imbalanced energy datasets.

2.4.3 Performance Metrics for Evaluating Models

Evaluating models designed to address class imbalance challenges in anomaly detection for energy consumption involves thoughtful consideration of metrics that capture the performance across normal and anomalous instances [143] In this context, several key metrics provide valuable insights into the effectiveness of these models:

2.4.3.1 Precision and Recall

Precision metric assesses the accuracy of the model when it predicts an instance as anomalous. A high precision indicates that the majority of instances predicted as anomalies are indeed anomalous. Recall gauges the model's ability to capture all actual anomalies. A high recall signifies that the model effectively identifies most of the true anomalies in the dataset [144].

2.4.3.2 F1 Score

The F1-score, also known as the harmonic mean of precision and recall, offers a balanced performance metric by considering both false positives and false negatives. It is particularly useful in scenarios where there is an imbalance between normal and anomalous instances[145].

2.4.3.3 Area Under the Receiver Operating Characteristic Curve (ROC-AUC)

ROC-AUC assesses the model's discriminatory ability between normal and anomalous instances across varying decision thresholds, with a higher AUC signifying superior discrimination. [146]

2.4.3.4 Area Under the Precision-Recall Curve (PR-AUC)

Similar to ROC-AUC, PR-AUC focuses on precision and recall. It is particularly informative in imbalanced datasets, offering insights into the trade-offs between precision and recall at various decision thresholds [146],[147].

2.4.3.5 Confusion Matrix

The confusion matrix offers a detailed tabulation of true positives, true negatives, false positives, and false negatives, serving as a basis for calculating other metrics and offering a granular understanding of model performance [1].

2.4.3.6 Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) puts into consideration all elements of the confusion matrix, providing a balanced measure of classification performance. When working with unbalanced datasets, it is quite helpful [148].

2.4.3.7. Balanced Accuracy

Balanced accuracy considers the imbalance in class distribution and calculates an accuracy score those accounts for both normal and anomalous instances. It provides a fair representation of overall model performance [80].

2.4.3.8 Precision-Recall at k (PR@k)

Precision-Recall at k evaluates precision and recall at a specific threshold (k). It helps understand the model's performance at different decision points[65], [144]

2.4.3.9 False Positive Rate (FPR) at Different Thresholds

Examining false positive rate at various decision thresholds provides insights into the model's ability to minimize false alarms while detecting anomalies [149].

2.4.3.10 Area Under the Cumulative Gain Curve (AUC-CG)

AUC-CG assesses the model's ability to rank anomalies higher than normal instances. It is particularly relevant when prioritizing the detection of critical anomalies [146].

In summary, a comprehensive evaluation of models dealing with class imbalance challenges in anomaly detection for energy consumption requires a combination of precision, recall, and other metrics that consider the unique characteristics of imbalanced datasets. These metrics collectively provide a nuanced understanding of

the models' performances, guiding the development of reliable and robust anomaly detection systems.

2.5 Ethical Considerations in Deep learning

The application of deep ensemble models for anomaly detection in energy consumption data raises significant ethical considerations, spanning privacy, fairness, transparency, and environmental impact. This review critically examines relevant literature to assess these ethical dimensions.

2.5.1 Privacy and Data Security

Energy consumption data, despite seeming innocuous, can disclose sensitive information about individuals' daily activities and lifestyles, raising substantial privacy concerns [150]. Granular data from smart meters, when analyzed, can infer occupancy patterns, appliance usage, and even sensitive personal information. Anonymization methods, such as data aggregation and differential privacy, are employed to mitigate these risks [151]. However, their limitations are evident; re-identification attacks, leveraging auxiliary data, can compromise anonymized datasets. Moreover, anomaly detection systems, if compromised, can expose vulnerabilities in the energy infrastructure, with potentially catastrophic consequences. Security implications extend to unauthorized access, the risk of data breaches, and manipulation of anomaly detection algorithms. Regulatory frameworks, for instance the General Data Protection Regulation (GDPR) and industry-specific standards, are crucial for enforcing data protection and accountability [152]. However, the dynamic nature of AI and energy technologies necessitates continuous adaptation of these frameworks.

2.5.2 Fairness and Bias

Anomaly detection algorithms, when trained on biased datasets, can perpetuate and deepen existing societal inequalities [153] Potential biases can arise from data collection processes, feature selection, and model training, leading to disparate impacts across different communities. For example, false positives in low-income communities might result in unwarranted service disruptions, while false negatives in affluent areas could lead to undetected energy theft. Methods for bias detection, such as fairness metrics and adversarial debiasing, are essential for mitigating these risks. Equitable distribution of benefits from advanced energy monitoring, including energy savings and improved grid reliability, is crucial for fostering social justice. Recent

studies have emphasized the need for intersectional fairness, considering the combined effects of multiple demographic factors [154] .

2.5.3 Transparency and Accountability

The ‘black-box’ nature of deep learning models postures challenges for transparency and accountability. Disclosure requirements for deployed AI systems are essential for ensuring that stakeholders understand the decision-making processes and potential impacts of these systems. Governance structures for algorithmic decision-making, including clear lines of responsibility and oversight, are crucial for preventing misuse and ensuring accountability. Audit procedures for anomaly detection models, including rigorous testing and validation, are necessary for detecting and correcting errors and biases. Stakeholder engagement in system development and deployment, including consultations with affected communities and experts, is crucial for building trust and ensuring that ethical considerations are adequately addressed. The need to provide clear explanations for detected anomalies is paramount, especially when anomalies lead to automated actions.

2.5.4 Environmental and Social Impact

The energy footprint of deep learning model training and deployment is a growing concern, particularly as models become larger and more complex [155]. The computational resources required for training and deployment of these models can contribute to greenhouse gas emissions and environmental degradation. Broader implications for energy conservation and sustainability, including the potential for AI-driven energy efficiency improvements, must be carefully considered. Social acceptance of AI-driven energy management, including public trust in the reliability and fairness of these systems, is crucial for their successful adoption. Ethical frameworks for responsible AI in critical infrastructure, for example the ‘IEEE Ethically Aligned Design and the European Commission's Ethics Guidelines for Trustworthy AI’, provide valuable guidance for ensuring that AI systems undergo responsible and ethical development and deployment [156]. Social impact assessments, including analyses of the potential effects on employment and community well-being, are essential for ensuring that AI-driven energy management contributes to sustainable development.

The ethical considerations surrounding deep ensemble anomaly detection in energy

consumption data are multifaceted and require a proactive and interdisciplinary approach. Future research should focus on developing robust and transparent AI systems that prioritize privacy, fairness, and accountability, while also considering the environmental and social impacts of these technologies.

2.6 Literature Review Gaps

A comprehensive review of the literature concerning models for dealing with class imbalance in energy consumption anomaly detection identified the following key research gaps:

1. **Limitations of Classical Methods:** Despite their prevalence, classical machine learning models like Random Forests and Support Vector Machines exhibit persistent challenges with class imbalance in this domain. Future research should prioritize: (a) enhancing SVMs for effective detection of sparse anomalies, particularly those distributed across the feature space, with a focus on capturing subtle rare anomaly patterns without exclusive reliance on margin maximization; and (b) developing Random Forest methodologies that improve the detection of sporadic rare anomalies through collective decision-making processes more attuned to their nuanced characteristics.

2. **Improvements in Deep Learning Ensembles:** While deep learning models (e.g., CNNs, LSTMs) represent the present cutting-edge for class imbalance in energy anomaly detection, their full potential remains untapped. A critical area for future investigation is the development and rigorous training of advanced deep learning ensemble models specifically designed to address this challenge effectively.

3. **Optimization of Training Strategies:** The prevalent use of minority class oversampling as a training strategy exhibits inherent limitations. Empirical research is warranted to evaluate and identify optimal training methodologies for diverse class-imbalanced energy datasets, potentially drawing upon successful techniques from other domains.

4. **Comprehensive Evaluation Framework:** The reliance on individual metrics (e.g., accuracy, AUC, F1-score, recall, precision) offers an incomplete assessment of model performance. Future empirical studies must adopt a multi-

metric evaluation approach to capture a holistic understanding of model capabilities. Furthermore, research is needed to systematically investigate the impact of various intrinsic dataset characteristics—including distribution (e.g., unimodal, multimodal), noise levels, imbalance ratio, temporal dependencies, feature space properties, specific anomaly types, data volume, and labeling quality—on the efficacy of different models and training strategies. This investigation is crucial for determining the most appropriate approaches for specific types of class-imbalanced energy consumption data.

2.7 Conceptual Framework

The conceptual framework is defined by its conceptual elements, the relationships between dependent and independent variables, and the operationalization of these variables [157]. A conceptual framework is aimed at helping the researcher to formulate important concepts and develop a theory to support the development of the resultant model for handling the class imbalance challenges in the anomaly detection in energy consumption [157]. In this research study, the conceptual elements, independent variables, moderating variables and dependent variables have been formulated as follows, in the subsequent sections.

2.7.1 Conceptual Elements

Understanding these conceptual elements helps in understanding the mechanics of the existing models for dealing with the problem of imbalanced class distributions in the context of anomaly detection for energy consumption. Moreover, it provides insights into the models' strengths and limitations, laying a strong foundation upon which new models are proposed.

2.7.1.1 Independent Variables

The independent variables are model architecture components and class imbalance handling techniques. Model architecture components include BiLSTM network configuration (number of hidden layers, hidden unit dimensions, dropout rates), CNN network configuration (filter sizes and numbers, pooling strategies, activation functions) and Ensemble integration method (voting mechanisms, weight allocation strategies). Class imbalance handling techniques, cost-sensitive learning parameters such as class weight ratios and misclassification costs.

2.7.1.2 Moderating Variables

Moderating variables are variables that act between the independent variable and dependent variable and change the strength of their relationship [157]. They provide a potential explanation or mechanism for how the independent variable influences the dependent variable [157]. In other words, moderating variables mediate or act as an intermediate step in the causal pathway between the predictor and the outcome [157]. In this research study, moderating variables included dataset characteristics, hyperparameter settings and energy consumption context. Data characteristics include class imbalance ratio (degree of imbalance (normal: anomaly ratio) and distribution of anomaly types), data quality factors (missing data percentage, noise levels, temporal consistency). Hyperparameter settings cover learning rates, batch size, training epochs and regularisation parameters. In the energy consumption context, the following are of interest: seasonal patterns (weather-related consumption variations, holiday/weekend effects), grid infrastructure (smart meter deployment density, grid modernization level) and consumer behaviour patterns (residential vs industrial consumption, peak usage periods). The technical environment entails computational resources (available processing power, memory constraints) and real-time requirements (latency tolerance, update frequency needs).

2.7.1.3 Dependent Variables

The dependent variable in this research was anomaly detection performance derived from the popular performance metrics for dealing with the class imbalance challenge in anomaly detection in energy consumption [158]. These metrics include the Confusion Matrix, Recall, Precision, Area Under the Precision-Recall Curve (AUC-PR), and F1 score. Figure 2.1 illustrates the relationship between the independent, moderating, and dependent variables.

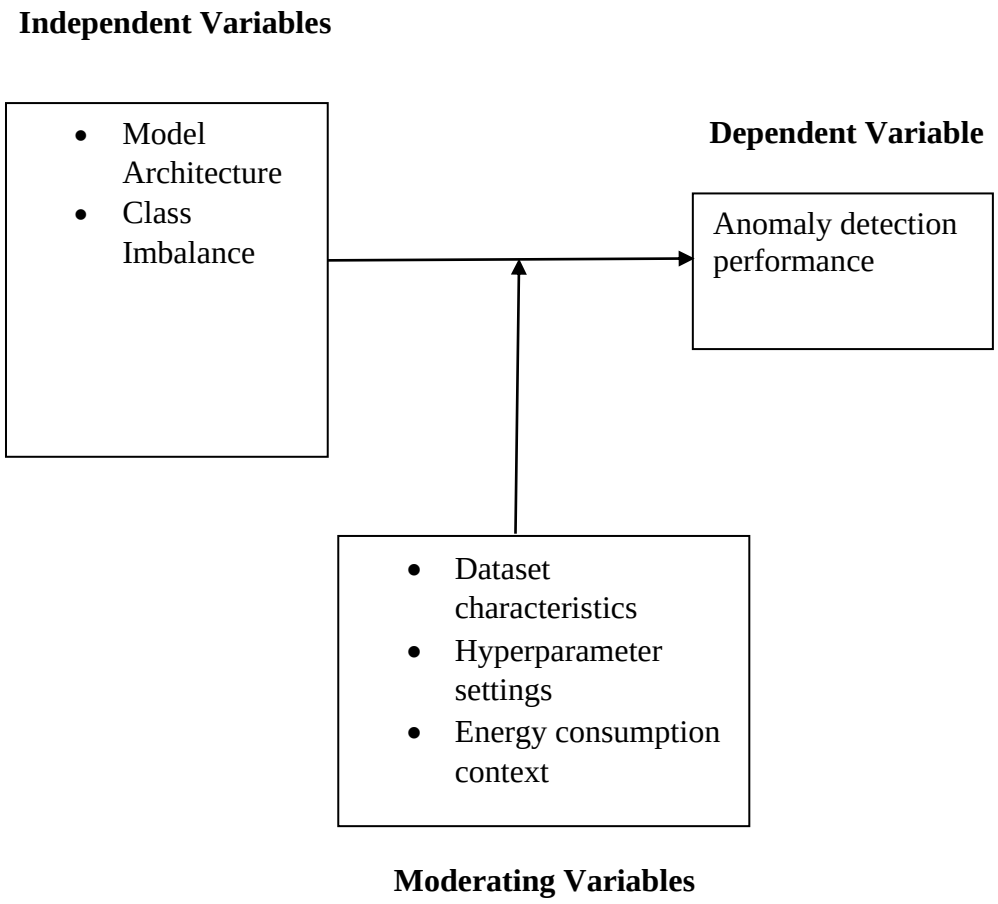


Figure 2.1: Conceptual Framework

2.7.2 Operationalization of the Variables

Variable operationalization involves defining and translating abstract constructs into measurable and observable indicators [159]. In research and data collection, operationalization is crucial as it allows researchers to quantify and gather empirical data to test hypotheses and draw meaningful conclusions [159]. Through the operationalization of variables, researchers can make the concepts of their study measurable, making them to put together empirical data that can be analyzed and interpreted with ease [159]. Clear operational definitions and valid measurement instruments are vital for warranting the accuracy and rigour of the research findings.

In light of this, this research uses the following steps in the operationalization of variables: identification of the constructs, operational definitions, selection of the measurement scales, development of the measurement instruments, pilot testing, data collection and data analysis. This approach is in line with the one proposed by Ryo and Rilling [159]. The constructs identified in this research include the specific scores

of the evaluation metrics from the different models in a variety of energy consumption data.

The independent and the moderating variables both use the nominal measurement scale while the dependent variables use the ratio measurement scales. Experiments were used to collect data on the operationalized variables. The initial experiment, in the form of a pilot study, was performed to demonstrate the feasibility of conducting this research study. In order to assess relationships between variables, test hypothesis, and draw conclusions based on the operationalized variables, data analysis was done using the R software and appropriate statistical techniques.

2.8 Chapter Summary

This literature review explored the multifaceted landscape of deep ensemble models for anomaly detection in imbalanced energy consumption data, highlighting the critical interplay between advanced machine learning techniques and the unique challenges posed by energy systems. The study reviewed various theories including the Extreme Value Theory, and Computational Learning Theory. Systems theory provided a foundational framework, emphasizing the interconnectedness and emergent behaviour within these complex adaptive systems, while statistical learning theory elucidates the mathematical underpinnings of model generalization and complexity. Deep learning architectures, including CNNs, RNNs, BiLSTMs, transformers, and GNNs, were reviewed that offer diverse capabilities for capturing spatial and temporal dependencies, necessitating careful selection and optimization to address the specific characteristics of energy consumption patterns.

The inherent class imbalance in anomaly detection, where normal consumption vastly outweighs anomalous events, necessitates the application of specialized handling techniques. Data-level approaches, including sampling methods, and algorithm-level approaches, like cost-sensitive learning, have been employed by researchers to mitigate this imbalance, each with its own strengths and limitations. Hybrid approaches and appropriate performance metrics, including precision, recall, accuracy, and AUC-ROC, are noted as crucial for robust model evaluation. Furthermore, explainability and interpretability methods, including LIME, SHAP, feature visualization, and attention-based techniques, are essential for building trust and understanding the decision-making processes of these complex models,

principally within critical energy infrastructure.

Ethical considerations, encompassing privacy, fairness, transparency, and environmental impact, are dominant in the development and deployment of AI-driven energy management systems. Privacy concerns arising from granular energy data necessitate robust anonymization techniques and regulatory frameworks. Potential biases in anomaly detection algorithms demand careful bias detection and mitigation strategies to ensure equitable distribution of benefits. Transparency and accountability are crucial for building trust, requiring clear disclosure requirements, governance structures, and audit procedures. Additionally, the energy footprint of deep learning models and their broader implications for sustainability must be carefully considered.

In summary, the development of a robust deep ensemble anomaly detection model for imbalanced energy consumption data requires a holistic approach, integrating advanced machine learning techniques with rigorous ethical considerations. Future studies may focus on developing adaptive, interpretable, and ethically sound models that contribute to the efficiency, reliability, and sustainability of modern energy systems.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter presents the research philosophy, research design and approach, population, sample and sampling procedure, data collection, data analysis and ethical considerations.

3.2 Research Philosophy

The study employed a pragmatic research philosophy, which emphasises the practical application of knowledge and the integration of diverse research approaches to address real-world problems. Pragmatism supports the use of both qualitative and quantitative techniques, aligning with the interdisciplinary nature of developing a deep ensemble model for energy systems, where technical efficiency and practical relevance are paramount. Pragmatism is rooted in the belief that the truth of a concept or method is determined by its practical consequences. In deep ensemble modelling, the ultimate goal is to improve prediction accuracy, robustness, and generalization by combining multiple models. This is a highly pragmatic approach, as the value of each model is judged not by theoretical elegance alone but by its real-world performance. Pragmatic philosophy is particularly well-suited for the development of deep ensemble models because it emphasizes practical outcomes, contextual understanding, and continuous learning, all of which align with the core principles and goals of ensemble learning in deep learning contexts. It encourages the exploration of diverse methods, prioritizes real-world effectiveness over theoretical purity, and embraces continuous adaptation and learning through experience. Pragmatism aligns with an experimental, trial-and-error approach. Ensemble development often involves testing different combinations of models, hyperparameters, and fusion strategies. The "best" ensemble is determined empirically.

Ontological Tenets

i. Objective Reality: Energy consumption patterns exist as measurable, objective phenomena independent of observer interpretation
ii. Reductionist Approach: Complex energy systems can be decomposed into analysable components and mathematical relationships
iii. Deterministic Causation: Anomalies result from identifiable, predictable factors

that can be computationally modelled.

Epistemological Tenets

i. Empirical Validation: Knowledge derived through systematic observation of energy consumption data and statistical testing
ii. Value-Free Inquiry: Research findings should be independent of researcher bias, relying on algorithmic objectivity
iii. Quantitative Measurement: Anomaly detection effectiveness measured through standardized metrics (precision, recall, F1-score, accuracy, AUC-ROC)
iv. Methodological Rigour: Ensemble model performance validated through cross-validation and multiple evaluation frameworks.

This study employed an experimental checklist to conduct experiments on the effectiveness of the current models used to deal with the problem of energy consumption class imbalance using a similar set of standard datasets. The evaluation of such models' performance, using a standard set of evaluation metrics, was recorded in a table of results before being analyzed.

The new deep ensemble model was developed based on the analysis findings of the experimental data from the main experiments. Based on the pragmatic research philosophy, the discussions and conclusions on the performance of both the newly developed model and the existing ones, in dealing with class imbalance issues in anomaly detection in energy consumption data, were based on the ground truth.

3.3 Research Design

This research implemented the mixed methods research design since there was collection and analysis of both the qualitative and quantitative data. Qualitative research design is a research approach that focuses on exploring and understanding the depth and nuances of a phenomenon [160]. Quantitative research design, on the other hand, is a research approach that is premised on systematic collection and analysis of numerical data so as to answer research questions, test hypotheses, and make statistical inferences [160]. In this research, the qualitative study analyzed literature on models for class imbalance in energy anomaly detection. The Khan et al. [161] five-step methodology was used during the literature review process. On the other hand, quantitative research design, entailed the collection and analysis of the

experimental data from the main experiments, through an experimental design methodology.

Using the design science methodology that incorporates the quantitative and the qualitative approach, this research aimed at solving the problem of addressing the problem of imbalanced class distributions in the context of anomaly detection for energy consumption, through the design of a new improved Deep Learning Ensemble Model based on weighted averaging. The research design included a strategy for achieving the objectives, mapping them to appropriate methods. This process ensured that the research process is both effective and efficient. The chosen methods are well-suited to address the research objectives while also taking into account both the practical constraints and ethical considerations within the research domain of addressing the imbalance issues in energy anomaly detection. This alignment ensured that the research study was focused, rigorous, and capable of generating valid and valuable insights. Table 3.1 shows how the research objectives and questions in this study were mapped to the respective research methods as proposed by Järvinen, [162].

Table 3.1: Strategies to Achieve the Specific Objectives

Objective	Method/Strategy	Description
To investigate existing models used to handle class imbalance in anomaly detection for energy consumption, with the aim of identifying effective techniques that can inform model selection for empirical analysis.	Desktop research design	The researcher reviewed journal articles and conference papers at this stage qualitative data was obtained
To conduct an empirical analysis of class imbalance handling models in the domain of anomaly detection for energy consumption	Experiment	Quantitative data was obtained from the State Grid Corporation of China (SGCC) dataset on which appropriate experiments were conducted
To develop a deep ensemble learning model that integrates cost sensitive learning, a class imbalance handling technique, for anomaly detection for energy consumption data	Design science	The researcher constructed an ensemble model from theoretical and empirical analysis.
To evaluate the performance of the newly developed deep learning ensemble model in addressing class imbalance challenge in the domain of anomaly detection for energy consumption.	Experiment	Quantitative data was obtained from the State Grid Corporation of China (SGCC) dataset on which appropriate experiments were conducted.

3.4 Experimental Design

The experimental research design in this study focused primarily on planning and conducting experiments, with an aim of investigating on the performance of the existing best-performing models (CNN, BiLSTM) for addressing the inherent challenge of skewed class representation in anomaly detection for energy consumption. The experiments also aimed at investigating on the most effective training process for the existing models for dealing with class imbalance issues in energy anomaly detection. In the empirical study, the existing best-performing models

used in the main experiments were identified through the theoretical analysis process. Based on the experimental design, research hypothesis was first formulated consistent with the research questions, before conducting the ablation study and the main experiments.

The Integrated Development Environment (IDE) for conducting the experiments was the Google Colab. The sampling strategy used in the selection of the anomaly detection datasets in energy consumption, models and the training strategies used in the various experiments are also described. Lastly, the experimental data collection process and techniques that was used are discussed, including the data analysis techniques for the experimental data. The proposal to adopt the design science approach in this research is aimed at enabling a problem-solving approach that combines elements of applied research, design thinking, and scientific inquiry to contribute both to the theoretical knowledge and practical applications in the field of anomaly detection for energy consumption.

3.4.1 Research Hypothesis

The following null hypothesis was formulated for the study:

H_0: "There is no statistically significant difference in F1-score, accuracy, precision, recall, or AUC-ROC between a deep ensemble model without cost-sensitive learning and other deep learning-based anomaly detection models on imbalanced energy consumption data".

In light of the research hypothesis, the experiments in the research study aimed at investigating the following:

1. If a deep ensemble learning model integrating cost-sensitive learning demonstrated statistically significant performance differences in energy consumption anomaly detection.
2. If a deep ensemble learning model without cost-sensitive learning exhibited statistically significant performance differences in energy consumption anomaly detection.

3.4.2 Pilot Study

A pilot study, within the domain of machine learning, denotes a limited-scope, initial

inquiry undertaken to evaluate the feasibility. and potential challenges of applying machine learning algorithms to a specific problem or dataset [163]. In this research study, a pilot study was conducted with the aim of assessing the feasibility of the main experiments as well as to guide on the selection process of the best performing models used in the main experiments. The challenges identified during the pilot study were used as the basis for improving the experimental design in the main experiments. Also, through the pilot study and the theoretical analysis, the experimental input datasets and the training processes used in main experiments were identified.

The pilot study provided an opportunity to adjust and fine tune both the experimental checklist and the observation checklist that were used as the main research instruments in this empirical study. During the pilot study, a cross-sectional survey on both the moderating variables and dependent variables of the existing models for dealing with the problem of imbalanced class distributions in the context of anomaly detection for energy consumption were investigated on purposively selected models that were recent and cutting-edge. The deep ensemble model was identified as the independent variables while the imbalance handling technique and energy consumption dataset characteristics as the moderating variables. The dependent variables are identified as the performance indicators of the selected models, measured using a set of evaluation metrics. Using the cross-sectional survey, an analytical approach was adopted to help focus on the cause and effect of the moderating variable on the dependent variable, derived from the performance indicators from the experimental models, as proposed by Buresh et al. [164].

A controlled experimental design was employed to incorporate controlled alterations in the deep learning setup and track their influence on a small scale. The control group was made up of the baseline model configuration and a set of default hyperparameters. The treatment included optimization algorithms, learning rate, and batch size.

3.4.3 Main Experiments

Conducting main experiments in this research study involved a systematic and rigorous process that first aims at identifying the independent variables, moderating variables and the dependent variables. In line with the research hypothesis, the existing best-performing models, class handling techniques, hyperparameter settings and data processing methods in energy consumption were identified as the independent

variables, data characteristics, energy consumption context, as well as technical environment as the moderating variables, while the primary performance metrics were identified as the dependent variables. The anomaly detection datasets in energy consumption used in the experiments were prepared through pre-processing in order to ensure quality, and split into both the training and the testing datasets.

The various models in the various experiments were trained and run in multiple times, recording the performance scores for each set of experiments, in a table of results. To identify the main effects and interactions among the different variables in the anomaly detection process from the class imbalanced datasets, the factorial design on these variables resulted in a number of experiments. In Table 3.2, the set of experiments, based on one of the best-performing models and anomaly detection datasets in energy consumption using class imbalance handling technique, are presented. For each experiment, the performance scores were recorded as the evaluation metrics.

Table 3.2: Experiments on the Various Models, Dataset and Class Imbalance Handling Technique

Model to be used	**Anomaly detection dataset in energy consumption**	**Class imbalance handling technique**
CNN	SGCC dataset	Oversampling the Minority Class (SMOTE)
CNN	SGCC dataset	Cost-Sensitive Learning
CNN	SGCC dataset	GAN
LSTM	SGCC dataset	Oversampling the Minority Class (SMOTE)
LSTM	SGCC dataset	Cost-Sensitive Learning
LSTM	SGCC dataset	GAN
BiLSTM	SGCC dataset	Oversampling the Minority Class (SMOTE)
BiLSTM	SGCC dataset	Cost-Sensitive Learning
BiLSTM	SGCC dataset	GAN
The ensemble model	SGCC dataset	Oversampling the Minority Class (SMOTE)
The ensemble model	SGCC dataset	Cost-Sensitive Learning
The ensemble model	SGCC dataset	GAN

3.4.4 Integrated Development Environment & Libraries

The Python IDE in Google Colab was employed to implement the various models used in the experiments. The choice of Google Colab over Jupyter Notebook was due to the former providing a free cloud-based environment giving access to Tensor Processing Unit (TPU) resources and Graphics Processing Unit (GPU) both advantageous for computationally intensive algorithms. Different Python libraries offer diverse functionalities. Scikit-learn, Imbalanced-Learn, XGBoost, LightGBM, TensorFlow, Keras Tuner, PyTorch, Resample-learn, and the Adversarial Robustness Toolbox are the primary libraries utilised for building, training, and evaluating. models on imbalanced datasets. Table 3.3 below shows the names of the different Python libraries that were used, including their specific uses.

Table 3.3: Python Libraries and Their Uses

Python Library	**Specific Use**
Scikit-learn	Scikit-learn provide a comprehensive set of tools for models used to handle class imbalance challenge.
Imbalanced-Learn	This library is an extension of Scikit-learn and specifically focus on addressing class imbalance through resampling techniques and cost-sensitive learning.
XGBoost	XGBoost is a powerful gradient boosting library that is used for tuning the hyperparameters with aim of dealing with the inherent challenge of skewed class representation in anomaly detection for energy consumption.
LightGBM	LightGBM is a gradient boosting framework used for efficient training on the models as it includes features to handle imbalanced datasets
TensorFlow and Keras Tuner	TensorFlow and Keras Tuner libraries were used for building and training deep learning models to address class imbalance challenge.
PyTorch	PyTorch is used for creating dynamic computational graphs when building neural network models.
Resample-learn	Resample-learn library extends Scikit-learn and imbalanced-learn functionalities by providing additional resampling techniques.
Adversarial-robustness-toolbox	This toolbox provides tools for assessing and improving the robustness of models for addressing class imbalance issues in energy anomaly detection, which was useful in addressing adversarial challenges in the anomaly detection process.

3.4.5 Research Instruments

Research instruments are specific tools, instruments, or materials that researchers use as part of a data collection technique. These instruments are designed to facilitate the collection of data accurately and systematically. Research instruments are often developed or adapted based on the research objectives and the chosen data collection technique. The choice of the research instrument plays a critical role in the quality, validity and reliability of the research findings [165]. In this research study, an experimental checklist, and an observation checklist were used as the main research instruments. These checklists ensure that the experiments on the performance of the different models were conducted systematically, consistently, and with attention to important details. They also help in effectively planning, executing, and documenting the results emanating from the several model experiments, in a variety of anomaly detection datasets in energy consumption and training processes. The observation checklist provides a mechanism through which the data from the various experiments was systematically recorded and documented to maintain consistency, objectivity, and thoroughness in the observations.

3.4.6 Research Population, Sampling and Sample Size

In this study, the target population is the publicly available SGCC dataset that displays 42,372 customers' daily electricity consumption in kilowatt-hours (kWh). The dataset was collected from January 1, 2014, to October 31, 2016 (1034 days). While 3,615 customers are classified as electricity thieves (labelled ‘1’), 38,757 customers are classified as regular electricity users (labelled ‘0’). Purposive sampling, often referred to as judgmental sampling, is a type of non-probability sampling in which the researcher selects study participants by using their professional judgment. Using expert knowledge of the population in making the decisions about the characteristics that are important to be represented in the sample is critical to purposive sampling[166].

This approach is in contrast to probabilistic sampling methods, such as random sampling, where each unit within the population has the same probability of being selected for the sample. In machine learning, purposive sampling can be helpful for both cost and time efficiency as opposed to large random samples that can be expensive and time-consuming [159]. In line with the purposive sampling strategy,

results of the theoretical analysis and the pilot study, the three best-performing models, benchmark anomaly detection datasets in energy consumption as well as the cutting-edge training processes were selected for use in the main experiments.

These have been presented in the subsequent Section i.e. experimental models, Experimental Datasets, Pre-processing and Training and Training Processes for the models. Benchmark datasets, in the context of machine learning and data analysis, constitute data employed for the standardized evaluation and comparative analysis of algorithmic performance. These datasets are widely accepted within a particular research community or field as representative of a specific problem or task [167]. They provide a common basis for assessing the performance of the machine learning models or algorithms. By using the same dataset, researchers and practitioners can objectively compare the effectiveness of different approaches [167]. Since benchmark datasets are publicly available and well-documented, they enable researchers to replicate experiments and results easily. This helps ensure the reliability and reproducibility of research findings. Lastly, benchmark datasets enhance the real-world relevance. Many benchmark datasets are chosen or designed to represent real-world problems or scenarios. This ensures that the evaluation of models on these datasets reflects their practical utility [167].

3.4.6.1 Experimental Models

Deep Neural Network architectures have been designed to solve problems in various domains or use cases. The choice of DNN architecture is informed by multiple factors, especially for classification problems[168]. DNNs offer three key advantages: first, their architecture enables the effective modelling of complex, non-linear dependencies inherent in energy consumption data. Second, they facilitate automated feature extraction from raw inputs. Third, DNNs frequently demonstrate superior, leading-edge performance in binary classification problems compared to conventional machine learning techniques like logistic regression and SVMs[168]. For purposes of this study, close attention is paid to CNNs, LSTM and BiLSTM models. The choice of these models was based on the promising results during both the theoretical analysis process and the pilot study. The newly developed model was known as the Deep Learning Ensemble Model for Handling Class Imbalance Challenge in the Anomaly Detection in Energy Consumption. This model was developed from the study of the

experimental data from the main experiments, besides the insights from the theoretical analysis.

3.4.6.2 Experimental Datasets, Pre-processing and Training

SGCC dataset was used as the experimental dataset in this research study. Table 3.4 shows the description of the high-dimensional benchmark dataset downloaded from the machine learning repositories.

Table 3.4: Description of the Experimental High-Dimensional Dataset

Experimental Dataset	Description	Number of records	Year of publication
State Grid Corporation of China (SGCC) dataset	The dataset displays 42,372 customers' daily electricity consumption in kilowatt-hours (kWh). While 3,615 customers are classified as electricity thieves (labelled '1'), 38,757 customers are classified as regular electricity users (labelled '0').	1,048,576 per meter	2016

Data pre-processing is an integral step in machine learning and entails pre-processing the raw data making it appropriate for modelling. Undeniably, a majority of real- world datasets are noisy, imperfect, and quite challenging for determining the behaviour of this data [14]. In line with this, the pre-processing of the primary data is very central in analyzing patterns from the dataset so as to realize accurate results. The raw data was transformed into numeric vectors. The energy consumption data comprises a variety of features; some of them are non-numeric types. To be able to use these attributes as input for the neural network, they were represented as numeric types. One-hot encoding was used to convert the non-numeric data into numerical ones. Data normalization aids in accelerating the solution, enhancing the models' accuracy, and preventing a feature with an exceptionally wide value range from interfering with the distance computation. The Min-Max Scaling function that maps raw data into a redefined interval, commonly set as [0, 1] which improves the performance of many machine learning algorithms by fetching features onto a similar scale. Min-Max

scaling stabilized gradient descent towards the global minimum [5] as shown in the following equation:

Xscaled = $(X - Xmin)/ (Xmax - Xmin)$

3.4.6.3 Training Processes for the Models

Oversampling the Minority Class, Under-sampling the Majority Class, Cost-Sensitive Learning, Ensemble Methods, Transfer Learning, Data Augmentation, Anomaly Generation, Custom Loss Functions, Adversarial Training as well as Sequential Learning and Online Training are common training strategies. Table 3.5 shows the description of the training strategies applied in deep learning. The outcomes of the experiments guided the decision-making on the new model's training process.

Table 3.5: Description of the class imbalance handling techniques

Experimental Training Methods	Description
Oversampling the Minority Class	This training strategy duplicates instances of the minority class to balance the class distribution[117]
Under-sampling the Majority Class	This training strategy randomly removes samples from the majority class to achieve a more balanced dataset [169]
Cost-Sensitive Learning	This training strategy allocates different misclassification costs to the minority and majority classes. It penalizes misclassifying the minority class more heavily to encourage the model to focus on accurately detecting anomalies [127]
Ensemble Methods dataset	This training strategy uses ensemble methods like Gradient Boosting or Random Forests with techniques that handle class imbalance well. It trains multiple models and combines their predictions to improve overall performance [169]
Transfer Learning	This training strategy leverages pre-trained models on larger, balanced datasets for related tasks and fine-tunes them on the imbalanced anomaly detection dataset[170]
Data Augmentation	This training strategy generates synthetic instances of the minority class to upscale its representation in the training dataset
Anomaly Generation	This training strategy creates synthetic anomalies to augment the minority class in the training data, ensuring a more balanced representation of anomalies [171]
Custom Loss Functions	This training strategy designs loss functions that give extra weight to misclassifications of the minority class [172]
Adversarial Training	This training strategy introduces adversarial examples during training to expand the model's robustness and sensitivity to anomalies [173]
Sequential Learning and Online Training	This training strategy trains the model sequentially, updating it as new data is rendered available. This is mostly beneficial for streaming data in energy consumption scenarios [174]

3.4.7 Data Collection Process and Techniques

Data collection techniques refer to the methods and procedures used to gather data or

information for a particular research study. These techniques are employed to collect primary or secondary data, or both, depending on the research goals and the nature of the study. Data collection techniques encompass a wide range of approaches, including surveys, interviews, observations, experiments, and questionnaires. The choice of data collection technique depends on the research objectives, the type of data needed, the target population, available resources, and ethical considerations [175]. Experiments were used to collect data. In the experiments, the performance of the different models in the anomaly detection dataset in energy consumption was recorded in a table of results.

3.4.8 Data Analysis Approach

The data analysis was done using the Jamovi software, and the aim was to compare the performance of the different models addressing imbalance issues in energy anomaly detection. The analysis was conducted using ANOVA on the performance metrics of the various models recorded in a table of results. ANOVA, or Analysis of Variance, is a statistical methodology employed to assess the presence of statistically significant differences among the means of two or more distinct groups within a given sample. ANOVA is normally used in hypothesis testing to determine whether there are statistically significant differences between the means of three or more independent groups [[176]. ANOVA aids in determining whether reported performance differences are statistically significant or the result of random variation. ANOVA is statistically justified for evaluating several models' performance, confirming ensemble improvements over baselines, and allowing organized post-hoc analysis to drive model modification. The metrics used in the models' evaluation process include the Confusion Matrix, Recall, Precision, F1 score, and Area Under the Receiver Operating Characteristic (ROC-AUC). The analysis results on the performance of the best existing three models (CNN, LSTM, BiLSTM) from the literature review were used as the foundation in the design and development of the newly developed Deep Learning Ensemble Model.

3.5 Model Development

Based on the data analysis results of the experimental data from the main experiments, the newly developed Deep Learning ensemble Model was designed and implemented using a design science research design approach. The proposal on the use of the deep

neural network in the new model was because of the fact from the literature review analysis, deep networks demonstrate flexibility in the hyperparameter settings during the training process. This would be a good leverage point in the development of a more effective model for dealing with the problem of imbalanced class distributions in the context of anomaly detection for energy consumption. Multi-methodological system methodology approach to this development composed of four strategies, namely: theory building, experimentation, observation and system development [177].

Theory building guides the system development process through the ideas contained in the conceptual framework [177]. Observation was achieved through surveys to provide important information within the research domain that guides both the experimentation process and the model development process [177]. In this research work, the experimentation process provides an avenue to experiment with the different models in anomaly detection. The experimentation process also provides an avenue to assess the effectiveness of the new model against the existing one. It also helps to validate the underlying theories in handling the class imbalance issues in anomaly energy detection. A model development process consisting of five stages, as proposed by Nunamaker *et al.*, [177] was adopted. These stages include: conceptual design, model architecture, model design, prototyping and experimentation. These have been explained in detail in the subsequent sections.

3.5.1 Application of Design Science Research in Deep Ensemble Model Development

This study adopts the Design Science Research (DSR) methodology to guide the development of a robust deep ensemble model for anomaly detection in energy consumption data. DSR offers a structured approach to constructing and evaluating artifacts intended to solve complex, real-world problems, making it well-suited for the domain of intelligent energy systems. The following six-step process was employed:

1. **Problem Identification and Motivation**

   The initial phase involved identifying the critical challenge of detecting anomalies within energy consumption datasets, which are often characterised by high dimensionality, temporal dependencies, and severe class imbalance. Such anomalies, though rare, can indicate inefficiencies, faults, or security breaches in power systems. The motivation stems from the need for intelligent models

capable of timely and accurate anomaly detection to enhance energy management and infrastructure resilience.

2. **Define the Objectives for a Solution**

The solution objectives were defined based on the problem context. The ensemble model was expected to: (i) accurately detect rare anomalous consumption events, (ii) maintain high recall and precision despite imbalanced classes, (iii) generalize well across different energy consumption patterns.

3. **Design and Development of the Artifact**

A deep ensemble model was designed as the core artifact. The model combines two deep learning architectures that is Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks. To address class imbalance, cost-sensitive training was integrated. The design ensured robustness, diversity among learners, and improved generalization capabilities.

4. **Demonstration**

The developed ensemble model was applied to publicly available and proprietary energy consumption State Grid Corporation of China (SGCC) benchmark dataset. These cases validated the model's applicability in operational contexts.

5. **Evaluation**

Model performance was rigorously evaluated using metrics appropriate for imbalanced classification, including Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Statistical tests such as the paired t-tests and ANOVA were employed to compare the deep ensemble model against baseline approaches and state-of-the-art deep ensemble models and determine significance in performance differences. The ensemble model consistently outperformed individual models.

6. **Communication**

The final step involved documenting the research findings, model design, and evaluation outcomes. This includes academic dissemination through peer-reviewed journal submission and the thesis.

By following the DSR framework, the study ensured a systematic approach to

model development, from conceptualization to practical validation.

3.5.2 Conceptual Design

This stage entails developing a conceptual framework. This framework assists the researcher in formulating key concepts and developing a theory to aid system development [177]. In this research study, the conceptual design consists of three variables that is independent variables, moderating variables and dependent variables. The model architecture components, class imbalance handling techniques, hyperparameter settings and data processing methods were used as the independent variables.

The data characteristics, energy consumption context and technical environment were used as the moderating variables while the primary performance metrics were used as the dependent variables in the conceptual design of the new model. Through the conceptual design, the blueprint of the new model was defined as follows: The energy consumption dataset was first input into the model. After this, the data was processed by handling both the missing data and the temporal aggregation. The feature engineering process and the training are then selected for model training. The model was finally evaluated and tested using a set of standard metrics. A conceptual model flowchart demonstrating the Ensemble model development flow is as shown in Figure 3.1.

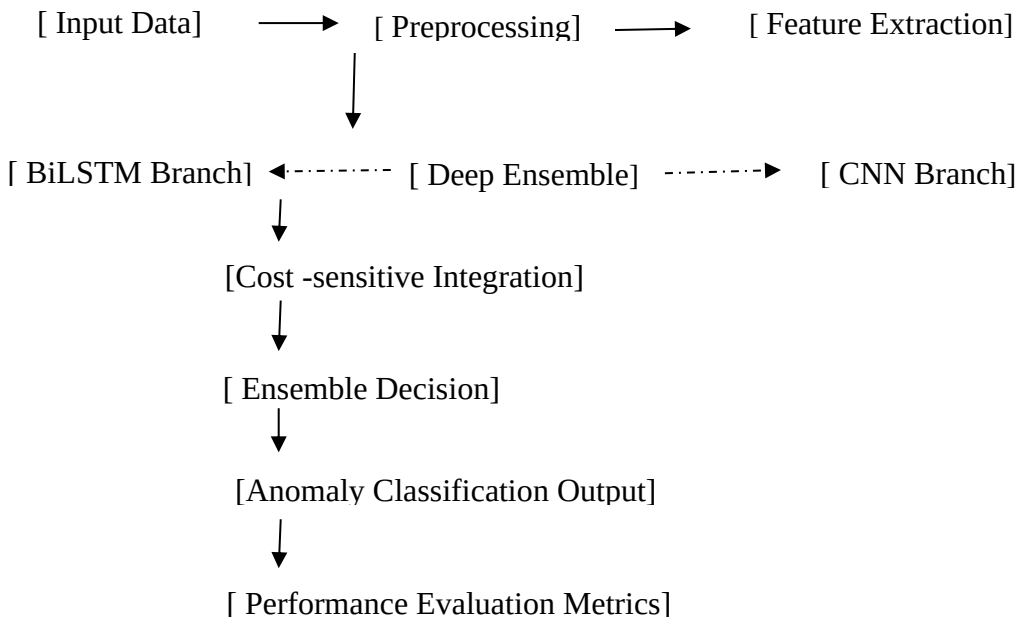


Figure 3.1: Conceptual Model Flowchart for Ensemble Model Development

3.5.3 Model Architecture

This stage entails the creation of a model architecture that serves as a blueprint for the model construction process. It also describes model components in terms of functionality, structural relationships, and dynamic interactions between them [177]. In this study, the energy consumption dataset was first input into the model. After the data pre-processing, the anomaly detection engine in the new model first performs feature engineering and a training process. The model was further evaluated and tested using a set of metrics. The intrinsic anomaly detection validation mechanism validates the quality of performance by comparing it with the model's output and the training-related hyperparameter settings.

The best training-related hyperparameter settings were picked at the model's best performance. Relatively low-performance scores trigger the deep model to perform further adjustments on the training-related hyperparameter settings to tune to better settings that improve the model's performance. The start of the self-adapting anomaly detection engine to the end of the stage where the best training-related hyperparameter settings are generated forms the top-level model, while the actual classification task by the model forms the baseline model. Figure 3.1 shows the architecture of the new Deep Learning Ensemble model for handling the class imbalance problem in anomaly detection in energy consumption.

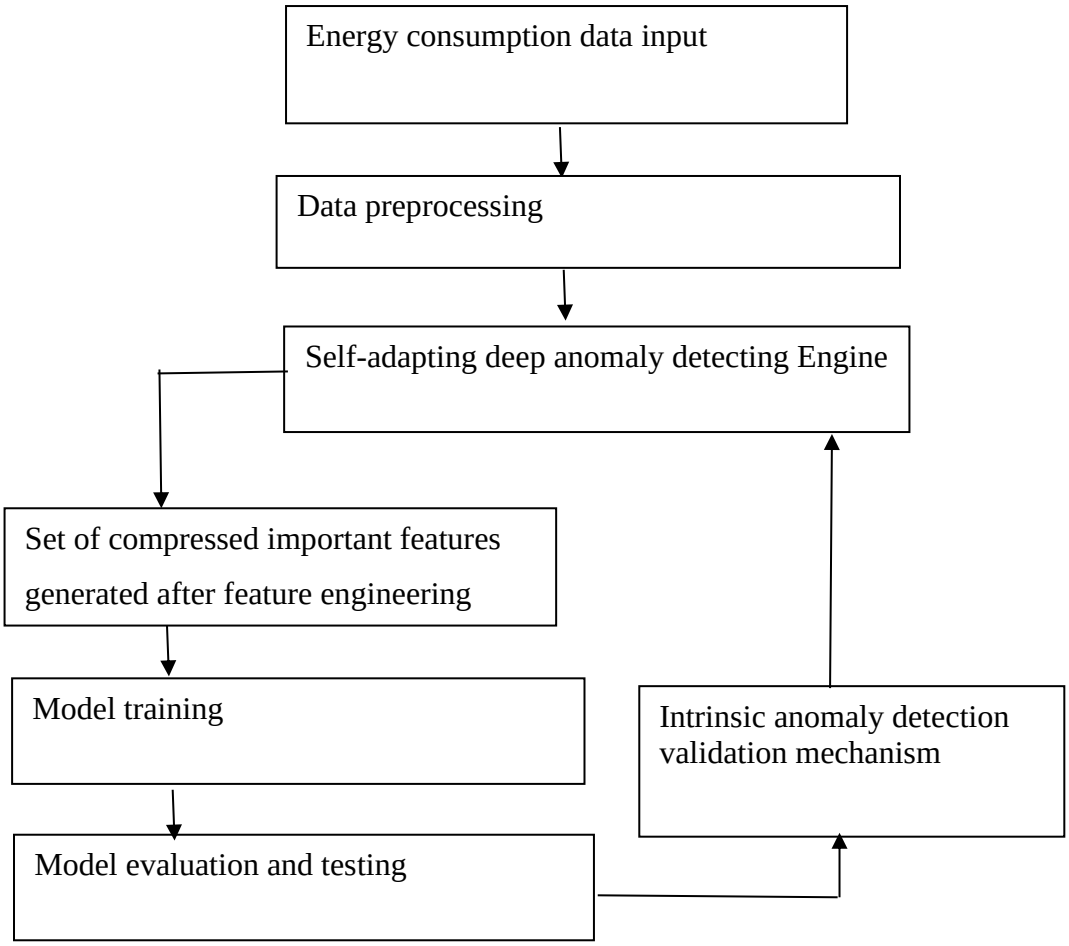


Figure 3.2: Architecture of the Deep Learning Ensemble Model for Addressing the Class Imbalance Challenge

3.5.4 Model Design

This phase involves model design and analysis. This involves understanding the area under study, using pertinent scientific and technological knowledge, developing a variety of alternatives, and synthesizing and assessing different solutions presented [177].In this research, the module design mainly focused on the setting up of detailed specifications for the model, including its architecture and its components. In the module design, the motivation was to create an effective model for handling the class imbalance problem in the anomaly detection of energy consumption data. The model design process mainly focused on the architecture and the interaction of different components of the self-adapting anomaly detection engine. The best-performing training process in the existing models (source task) was used to perform supervised transfer learning on the training process in the new model (target task), making it more effective in the training process and subsequently in addressing the class imbalance

problem in anomaly detection in energy consumption data.

The best-performing existing models (CNN, BiLSTM) were trained on the source task, using the best-performing training process on the existing models. The knowledge captured by the existing model using this best-performing training process was extracted in the form of learned model weights. These learned model weights were then used to perform transfer learning initialization into the training process of the new model (or the target task). Fine-tuning process using the new model's training process continued iteratively until the model's performance on the target task was optimal. Transfer Learning (TL) was chosen over alternative paradigms such as Meta-Learning and Hybridization in developing a deep ensemble model for anomaly detection in energy use due to empirical fit, data availability, complexity, and computational practicality.

3.5.5 Prototyping

In this phase, a prototype model is created to test the new model in an actual environment. This prototype can be further developed into a finished product and integrated into a fully functional system [177].

In this research, the rapid application development methodology was adopted to create the model's prototype to assess the technical feasibility of the newly developed model. This prototype also provides an avenue for further improvements on the model, especially on its training process.

3.5.6 Experiments

This stage involves the evaluation process through experiments. This aids in the researchers' observation of the model's functioning, performance, and the impacts of the model [177]. In this research, different experiments were used to advance on the performance of the new model through the improvement of its training process. First, experiments were done on the performance of the different training processes in a variety of models and energy consumption dataset. The best training process was used to perform the supervised transfer learning process on the new model learning. Figure 3.3 gives an illustration of the five stages together with a description of how these stages were adopted to fit the development of the new model. The model development methodology addresses the five stages followed in model development.

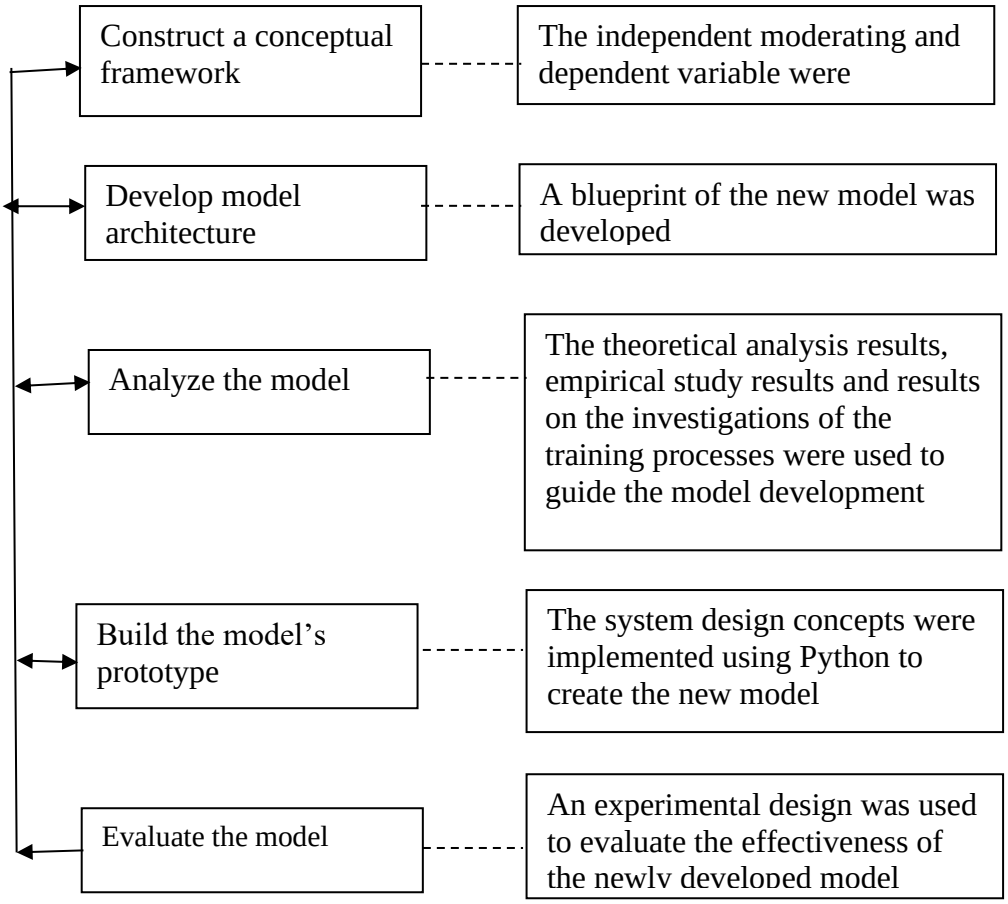


Figure 3.3: Research methodology for system development

3.6 Ethical Considerations

The researcher obtained ethical clearance from the Directorate of Postgraduate Studies at Murang'a University of Technology before commencing the study. Subsequently, a research permit was secured from the National Commission for Science, Technology and Innovation in Kenya, in accordance with national research regulations. In this research, the guidelines associated with copyright, authorship, and patenting policies, including data-sharing policies were closely adhered to.

3.7 Chapter Summary

The study employed a pragmatic research philosophy, which emphasizes the practical application of knowledge and the integration of diverse research approaches to address real-world problems. A Design Science Research (DSR) approach was employed to guide the creation and validation of the deep ensemble model. This design enabled the iterative refinement of the model to ensure its applicability in real-world energy consumption environments. The experimental framework was based on a comparative evaluation of multiple deep learning models. The deep learning models were trained

and tested on a time-series energy consumption dataset (SGCC) exhibiting class imbalance. Experiments were designed to compare the performance of the ensemble model with individual baseline models using repeated k-fold cross-validation to enhance robustness. Hyperparameters were tuned using grid search and early stopping to avoid overfitting. Quantitative analysis was performed to evaluate the model's performance across different scenarios. Statistical methods were used to determine whether observed differences in performance were significant. Evaluation focused on both overall accuracy and anomaly detection capability, emphasizing recall and AUC due to the imbalanced nature of the data. Ethical protocols were strictly observed throughout the study. All datasets used were either publicly available or anonymized to ensure data privacy.

CHAPTER FOUR

ANALYSIS OF MODELS FOR ANOMALY DETECTION BASED ON CLASS IMBALANCE

4.1 Introduction

This chapter helps to address research objective one of the study, which was to "review existing models used to address class imbalance challenge in the domain of anomaly detection for energy consumption"

Energy consumption monitoring and management have undergone a significant transformation with the spread of smart grids and advanced metering infrastructure [29]. The availability of high-resolution energy consumption data provides a unique opportunity to enhance anomaly detection, aiding the detection of infrequent patterns that significantly depart from normal usage patterns. These anomalies can manifest in various forms, including sudden load changes, equipment malfunctions, unexpected fluctuations in energy consumption, or potential cyberattacks.

The importance of accurate and timely anomaly detection cannot be overstated, as it serves several critical purposes. In terms of energy efficiency, anomaly detection helps identify and address inefficiencies, equipment failures, and energy waste, leading to substantial reductions in energy consumption and associated costs. Grid stability also benefits significantly from anomaly detection, as rapid identification of sudden load changes or equipment failures is essential for maintaining grid stability and preventing cascading failures that could disrupt power supply.

Operational reliability is another crucial aspect enhanced by anomaly detection. By identifying and addressing irregularities, it enables proactive maintenance and repair of critical equipment, minimizing service disruptions and ensuring consistent energy supply. Furthermore, anomaly detection is pivotal in cybersecurity, serving as a critical tool for detecting and countering cyberattacks targeting energy infrastructure, thereby protecting crucial resources and maintaining the integrity of the grid.

However, the intrinsic class imbalance presents a major obstacle to creating efficient anomaly detection models for energy consumption. By definition, anomalies are infrequent occurrences. As a result, the dataset becomes skewed, with anomalous

events greatly outnumbered by typical patterns of energy consumption. As a result of this class imbalance, machine learning algorithms may be biased towards normal data, making it difficult for them to detect uncommon but crucial anomalies.

This chapter synthesizes and critically evaluates pertinent literature to clarify the strengths, weaknesses, and future trajectories of existing techniques, encompassing both advanced deep learning architectures and traditional machine learning methods, with a focus on their efficacy in managing class imbalance and their adaptability to real-world energy data.

4.2 Review of Existing Models

Anomaly detection is a critical task in machine learning that involves identifying patterns in data that do not conform to expected behaviour. These anomalous patterns often correspond to significant, rare, or potentially critical events. The review was thematically carried out.

4.2.1 Traditional Anomaly Detection Methods

Conventional anomaly detection methods for energy consumption analysis are broadly classified into statistical and machine learning approaches. Statistical methods form the foundation of anomaly detection, relying on statistical models to identify deviations from expected patterns. These include techniques such as thresholding, statistical process control (SPC), and time series analysis, for instance, Autoregressive Integrated Moving Average (ARIMA) [178] models. While widely used, these statistical approaches have limitations in their capacity to model complex, non-linear relationships within energy consumption data and may be particularly sensitive to noise and outliers.

Machine learning methods offer alternative approaches to anomaly detection, with SVMs [179] being a prominent example. SVMs function by identifying an optimal hyperplane that maximizes the margin between different classes in the feature space thus separating normal data points from anomalies. However, they encounter challenges when subjected to imbalanced datasets, they become biased towards the majority class, leading to suboptimal decision boundaries and reduced accuracy in detecting rare anomalies [137].

K-Means clustering is another machine learning approach that uses an unsupervised

learning algorithm to group related data points. In this context, anomalies are identified as data points that fall outside of dense clusters. The method's effectiveness is limited by the challenge of determining the ideal number of clusters (k), and it may struggle to handle complex, non-linear data distributions effectively. Additionally, K-Means might not accurately identify anomalies that manifest as more than simple outliers.

The Isolation Forest method takes an ensemble-based approach to anomaly detection, constructing multiple isolation trees to identify unusual data points [22]. This method considers data points that are easily isolated from the other items of the dataset as anomalies. While Isolation Forest can be effective in certain scenarios, it presents challenges in hyperparameter tuning, and its performance can vary significantly depending on specific characteristics of data and the degree of class imbalance present.

4.2.2 Deep Learning Models

Deep learning models have emerged as promising solutions for anomaly detection, distinguished by their capability to automatically learn complex features from data. Convolutional Neural Networks (CNNs) [180] have been particularly effective in extracting spatial features from data, enabling identification of anomalies based on spatial patterns in energy consumption. These patterns might include unusual variations across different time periods or locations, making CNNs valuable tools for comprehensive anomaly analysis.

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) network [181], excel in analyzing sequential data such as time series energy consumption patterns. Their architecture is specifically designed to capture temporal dependencies, allowing identification of anomalies by recognizing deviations from established temporal patterns and trends. This makes them especially valuable in energy consumption monitoring where temporal relationships are crucial.

Autoencoders [25] represent another powerful approach in the deep learning toolkit. The training objective for unsupervised learning models involves the reconstruction of input data, offering a unique method for anomaly detection. The key principle lies in measuring reconstruction error – when the model encounters anomalous data, it typically produces larger reconstruction errors, providing a natural mechanism for

identifying unusual patterns.

4.2.3 Techniques for Handling Class Imbalance

The challenge of class imbalance in anomaly detection has yielded to the development of several sophisticated mitigation techniques. Oversampling methods include Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic minority class samples by interpolating between existing minority class instances [31]. This method improves the model's ability to learn from rare events by increasing minority class representation. Similarly, Adaptive Synthetic Sampling (ADASYN) generates synthetic samples in regions where the minority class is underrepresented, strategically focusing on areas where additional samples are most beneficial [135].

Class imbalance can be addressed through various undersampling approaches, ranging from simple to sophisticated. The most basic method attains a balanced class distribution by employing a random undersampling technique on the majority class. In contrast, advanced techniques like NearMiss take a more nuanced approach, strategically removing the majority class instances that lie nearest to minority class samples. This method preserves the most relevant information from the majority class while still achieving the desired reduction in representation.

An alternative strategy for handling class imbalance comes in the form of cost-sensitive learning. This method implements a sophisticated weighting mechanism during the training process, assigning higher weights to minority class instances. By doing so, the model becomes more attuned to identifying rare anomalies, even though this might lead to increased errors in majority class classification. This trade-off proves exclusively valuable in circumstances where failing to detect anomalies could have serious repercussions.

4.3 Methodology

This research employed a systematic literature review methodology utilizing the five-step framework advanced by Khan, et al. [161]. This framework provides a structured approach for conducting a comprehensive and unbiased literature review.

i. **Framing Research Questions:**
   a. **RQ1:** What models are used to address the class imbalance challenge in anomaly detection in energy consumption?
   b. **RQ2:** What are the strengths and limitations of existing models in handling class imbalance in this domain?
   c. **RQ3:** What are the key challenges and research gaps in current anomaly detection approaches for imbalanced energy consumption data?

ii. **Identifying Relevant Literature:** A detailed search was conducted using relevant databases, including ACM Digital Library, Google Scholar, IEEE Xplore, Web of Science, and Scopus. Key search terms included "anomaly detection," "energy consumption," "class imbalance," "deep learning," "SVM," "K-Means," "isolation forest," "oversampling," "undersampling," "cost-sensitive learning," and combinations thereof.

iii. **Assessing Article Quality:** Articles were selected based on relevance, quality, and recency. Inclusion criteria included:
   a. Published within the last five years (between 2019 and 2023) to ensure the inclusion of recent advancements.
   b. Focus on anomaly detection in energy consumption data.
   c. Explicit discussion of class imbalance handling techniques.
   d. Availability of sufficient information regarding methodology, results, and conclusions.
   e. Published in reputable academic journals or conferences.

iv. **Conducting a Critical Review:** A thorough analysis of the selected articles was conducted, focusing on:
   a. The specific anomaly detection methods employed.
   b. The techniques used to address class imbalance.
   c. The performance evaluation metrics used (recall, accuracy, F1-score, precision, AUC-ROC, etc.).
   d. The limitations and challenges identified in each study.
   e. The data sources and experimental setups used in each study.

v. **Interpreting Results:** The findings were synthesized and interpreted to highlight key trends, challenges, and promising research directions in the field of anomaly detection for imbalanced energy consumption data.

4.4 Results and Discussion

4.4.1 Classical Machine Learning Models

In the literature, SVM [16] represent a foundational approach in machine learning-based anomaly detection, operating by identifying optimal hyperplanes to isolate different classes in the feature space. While SVMs have demonstrated considerable success in various classification tasks, they face significant challenges when working with imbalanced datasets typical in anomaly detection situations [16]. Their core mechanism of margin maximization, while mathematically elegant, can cause biased decision boundaries that favour the majority class. This bias becomes particularly problematic when anomalies are sparsely distributed across the feature space, as the algorithm may struggle to establish appropriate decision boundaries that can effectively capture these rare instances[19].

Random Forests, despite leveraging the power of ensemble learning through multiple decision trees, encounter similar challenges when confronted with imbalanced datasets. The fundamental issue lies in their training process, where each tree is constructed using bootstrap samples from the training data[58]. In imbalanced datasets, these samples naturally contain more instances from the majority class, leading to trees that are inherently biased towards majority class predictions. Bias manifests in reduced sensitivity to anomalies, as the collective voting mechanism tends to favour the predominant class. Even though Random Forests offer advantages like feature importance ranking and robust performance across various domains, their effectiveness in anomaly detection can be compromised when the class distribution is heavily skewed [58]. This limitation becomes particularly apparent in scenarios where the accurate identification of rare anomalies is crucial, such as in energy consumption monitoring or fault detection systems.

This inherent bias in both SVM and Random Forest models highlights the importance of implementing appropriate preprocessing techniques or modifying the base algorithms to better handle class imbalance. Such modifications might include adjusting class weights, implementing specialized sampling techniques, or developing

hybrid approaches that integrates the strengths of multiple methods to achieve more balanced and effective anomaly detection.

4.4.2 Deep Learning Models

The evolution of deep learning models, particularly CNNs and RNNs, including LSTMs, has marked a significant advancement in anomaly detection capabilities [14],[182],[60]. These sophisticated architectures have exhibited remarkable potential in identifying and analysing complex patterns within data. Notable research in this field includes the work of Huang, et al [183], who developed a deep learning-based model specifically designed for smart grid anomaly detection, achieving significant improvements in both accuracy and precision. Similarly, Gao et al[184]. contributed to the field with their deep autoencoder-based model for industrial process anomaly detection, successfully demonstrating its capability to recognise subtle deviations from normal operating circumstances. These studies serve as important benchmarks for evaluating new approaches in this field.

Despite their promising capabilities, deep learning models face several significant challenges in anomaly detection. A primary concern is their ability to effectively detect sparse anomalies dispersed across the feature space, particularly in the case of CNNs[60]. Additionally, these models often struggle with class imbalance, potentially developing biases that favour the majority class at the expense of detecting rare but critical anomalies.

4.4.3 Limitations of Existing Approaches

The limitations of existing approaches extend across various domains. The handling of class imbalance remains a persistent challenge, with many studies relying heavily on oversampling techniques such as SMOTE [31]. While these methods can be effective, they risk introducing artificial data points that may not accurately represent true anomaly distributions, potentially leading to overfitting, especially in data-limited scenarios.

Each model type exhibits specific limitations. SVMs struggle with hyperparameter sensitivity, particularly in kernel function and regularization parameter selection. K-Means clustering faces challenges in optimal cluster number determination and handling non-linear data distributions. Isolation Forests show sensitivity to

hyperparameter tuning and may contend with complex, multi-modal anomaly patterns. Deep learning models, while powerful, face challenges in capturing sparse anomalies and managing complex temporal dependencies, alongside the persistent problem of class imbalance during training.

The evaluation methodology of these approaches also requires careful consideration. While traditional metrics like precision, accuracy, F1-score and recall are commonly employed, a more inclusive evaluation framework is necessary. This should incorporate metrics like AUC-ROC, precision-recall curves, and cost-sensitive metrics that explicitly account for class imbalance. Furthermore, many studies focus on specific datasets, potentially limiting their generalizability to diverse energy consumption scenarios with varying characteristics, consumer types, noise levels, and anomaly patterns.

4.5 Key Insights

Several key insights emerge from this analysis. Class imbalance stands out as a critical challenge that significantly impacts model performance in energy consumption anomaly detection. Deep learning models, despite their limitations, offer substantial potential for improving detection accuracy, particularly in capturing complex patterns and temporal dependencies. The development of effective class imbalance handling techniques remains crucial for ensuring system robustness and reliability. Finally, the implementation of a comprehensive evaluation methodology is essential, considering various performance metrics while dealing with the specific challenges posed by class imbalance in real-world applications.

Deep learning models, mainly CNNs and RNNs (particularly, LSTMs), have shown promise in capturing complex patterns and increasing anomaly detection performance.

4.6 Discussion

The literature review reveals several crucial challenges and promising directions for future research in anomaly detection within imbalanced energy consumption data [16],[20]. The examination of existing models has highlighted significant limitations across both classical and modern approaches. Classical machine learning models, including SVM [16] and K-Means [20], though theoretically robust, demonstrate consistent difficulties in managing the inherent class imbalance in energy consumption

data. SVMs, despite their strong theoretical underpinning, often produce biased decision boundaries when confronted with imbalanced datasets, leading to reduced accuracy in anomaly detection [18]. K-Means clustering, while effective for identifying data point clusters, shows sensitivity to initial cluster selection and struggles with anomaly identification beyond simple outliers.

Deep learning approaches, despite their advanced capabilities, face their own set of challenges. Autoencoders, for instance, can be particularly vulnerable to class imbalance issues, potentially producing biased reconstructions that compromise their ability to accurately identify rare anomalies [9]. This highlights the critical need for developing effective training strategies specifically designed to address class imbalance in deep learning architectures.

The investigation of more intricate class imbalance handling techniques emerges as a crucial research direction. While established methods like SMOTE [31],[76] have demonstrated some effectiveness, their application in complex scenarios often reveals limitations. Future works ought to focus on developing hybrid approaches that combine various sampling techniques, and investigate the potential of generative adversarial networks (GANs) for creating more diverse and representative synthetic minority class samples.

The integration of external data sources presents a significant opportunity for enhancing model performance. Energy consumption patterns are influenced by numerous external factors, including weather conditions, economic indicators, and social media trends. Incorporating these diverse data sources into anomaly detection models could provide essential context and improve the capture of complex consumption patterns.

The development of explainable AI (XAI) models represents another critical research direction. Model interpretability is crucial for trust and effective human-in-the-loop decision-making. This transparency becomes particularly crucial when investigating the root causes of anomalies and ensuring the reliability of detection systems.

Real-time implementation of anomaly detection models stands as a vital consideration for practical applications. The capability to detect and react to anomalies in real time

enables proactive measures against equipment failures, cyberattacks, and other potential threats to grid stability. This capability is paramount to maintaining the reliability and security of energy infrastructure.

Further, the persistent challenge of minority class representation has been a critical limitation in traditional anomaly detection methodologies. Researchers have progressively shifted to sophisticated ensemble learning techniques and generative models to mitigate this fundamental issue. Ahmed et al. [34] study established the potential of Deep Ensemble Learning Models (DELM) to overcome data skewness by employing techniques such as Conditional Generative Network to augment minority class data. Similarly, the research by [35] utilizing Conditional Generative Adversarial Networks (CGAN) highlights the importance of advanced representation learning for distinguishing rare anomalies within predominantly normal datasets.

The implications of these approaches extend beyond mere technical solutions, fundamentally transforming our understanding of anomaly detection in imbalanced datasets. By integrating diverse neural network architectures, preprocessing techniques like batch normalization, and advanced feature extraction methods, researchers are developing more adaptive and context-aware anomaly detection frameworks. The proposed research objective of developing a deep ensemble model directly addresses critical computational theory gaps, seeking to balance detection accuracy with computational efficiency. Techniques such as the ‘selective ensemble method for anomaly detection based on parallel learning (SEAD-PL)’ demonstrate the potential for creating flexible, dynamic frameworks that can adjust to varying data distributions and capture subtle consumption irregularities [38]. This solution not only augments the technical capabilities of anomaly detection but also offers a more nuanced understanding of complex energy consumption patterns, bridging the gap between computational theory and practical energy management strategies.

These findings institute a strong foundation for subsequent research chapters, informing the development and evaluation of novel deep ensemble learning approaches. Future models ought to address existing limitations by incorporating diverse deep learning architectures, implementing effective class imbalance handling techniques, and leveraging the comprehensive insights gained from this literature review.

4.7 Chapter Summary

This chapter presented an analysis of existing anomaly detection models for energy consumption data, placing particular emphasis on dealing with the fundamental challenge of class imbalance. The review encompassed a broad spectrum of methodologies, ranging from established machine learning approaches like SVMs and K-Means clustering to advanced deep learning architectures, including CNNs and RNNs.

A critical finding emerged from this analysis: class imbalance presents a persistent challenge that significantly impacts model performance across different approaches. Traditional machine learning models, despite their proven effectiveness in certain contexts, frequently struggle with accurate anomaly detection due to inherent biases favouring the majority class. While deep learning models demonstrate promising capabilities, they too face similar challenges when confronted with rare and complex anomalies, particularly in highly imbalanced datasets.

The review emphasized the crucial role of effective class imbalance handling techniques, including under sampling, oversampling, and cost-sensitive learning strategies. However, the investigation revealed that the success of these approaches varies considerably depending on specific data characteristics and model architectures, highlighting the need for context-specific solutions.

Several significant gaps in current research were identified through this comprehensive review. The shortcomings of classical machine learning models became apparent, particularly in their ability to deal with complex, imbalanced datasets. Similarly, the challenges associated with training and evaluating deep learning models in imbalanced scenarios were thoroughly documented. The review highlighted the pressing need for more sophisticated class imbalance handling techniques, suggesting the exploration of hybrid approaches that combine various sampling methods and the potential utilization of GANs for creating synthetic data.

The incorporation of external data sources emerged as a promising direction for future research, with potential benefits from incorporating weather data, economic indicators, and other relevant external factors. The development of explainable AI models was identified as crucial for building user trust and enabling effective human-

in-the-loop decision-making processes. Additionally, the review emphasized the importance of developing and implementing real-time anomaly detection systems to ensure timely response capabilities and operational effectiveness.

These findings provide essential insights into challenges and opportunities present in anomaly detection for imbalanced energy consumption data. The insights gained from the literature review acted as a foundation for the development and evaluation of novel deep ensemble learning approaches proposed in subsequent chapters.

CHAPTER FIVE

EMPIRICAL ANALYSIS OF CLASS IMBALANCE HANDLING MODELS

5.1 Introduction

An experimental study helped to address research objective two, which was “to conduct an empirical analysis of class imbalance handling models in the domain of anomaly detection for energy consumption “.

Accurate modelling and analysis are severely hampered by the prevalence of class imbalance in data, especially in the area of anomaly detection. Anomalies, which frequently indicate odd increases or decrease in usage, are generally uncommon in energy consumption datasets when contrasted with typical consumption trends. Models with a predisposition towards the majority class may result from this imbalance, which eventually reduces their ability to detect important anomalies that might point to underlying problems like equipment malfunctions or fraudulent activity.

The need to assess and contrast current class imbalance handling models designed especially for energy consumption data necessitates an empirical investigation. Although many methods have been put forth in the literature, the context and type of dataset can significantly affect how effective they are in practice. This study intended to provide a data-driven basis for future model development by carrying out an extensive empirical analysis in order to reveal insights into which approaches function best under particular circumstances.

This empirical study focused on several well-known strategies that have been utilized to address class imbalance issues in anomaly detection. Among the techniques chosen were the Synthetic Minority Over-Sampling Technique (SMOTE), a technique that balances the dataset by generating synthetic samples for the minority class [27]. Cost-sensitive Learning, which adds misclassification costs, modifies the learning algorithm to give the minority class more consideration. This study also used the Generative Adversarial Networks (GAN). These techniques were selected due to their widespread use in earlier studies and their theoretical applicability for managing unbalanced datasets in the context of anomalies in energy consumption.

The objective of the empirical analysis was to present a thorough assessment of these techniques using cutting-edge deep learning models by evaluating their performance using recognized metrics such as precision, accuracy, F1-score, and recall. In order to guide the development of a more effective deep ensemble model for anomaly detection in imbalanced energy consumption data, the findings helped clarify the advantages and disadvantages of each strategy.

5.2 Methodology

This section provides an in-depth examination of deep learning models used in conjunction with class imbalance handling techniques to discover energy usage anomalies. The study focused on how well various architectural approaches and data balancing techniques perform in comparison.

5.2.1 Data Collection and Preprocessing

The SGCC dataset, which offers thorough electricity consumption data labelled for anomaly detection, was used for the empirical analysis. The dataset displays 42,372 customers' daily electricity consumption in kilowatt-hours (kWh). The data was collected from “January 1, 2014, to October 31, 2016 (1034 days)”. While 3,615 customers are classified as electricity thieves (labelled 1), 38,757 are classified as regular electricity users (labelled 0). The SGCC (State Grid Corporation of China) energy consumption dataset was chosen for its accessibility, anonymized as well as its ability to fill a substantial research gap in the field of anomaly detection in smart grid systems. The dataset is publicly available, well-documented, and formatted for time-series analysis, making it an ideal choice for reproducible research. Because of its accessibility, other researchers can replicate and extend the work, complying with open science standards and promoting cross-study benchmarking. As a result, the SGCC dataset was chosen as a primary data source for model construction and evaluation because of its availability and representativeness of real-world energy usage patterns.

Normal consumption patterns are designated as 0 and anomalous patterns as 1 in the dataset's binary classification system. This methodical approach makes it possible to distinguish between typical usage patterns and possible theft or equipment failures. Data preprocessing included preparing sequences that were appropriate for deep learning architectures, extracting temporal features, and normalizing consumption

values. As part of the data preparation process, the SGCC dataset was split into training and test sets using an 80:20 split ratio. The 80:20 data split ratio was chosen to ensure sufficient training data for robust learning, enable meaningful and unbiased evaluation, address class imbalance in anomaly detection, and align with widely accepted standards in machine learning. A smaller training portion could lead to underfitting, particularly when dealing with high-dimensional time-series data. This split is commonly referenced in the literature and provides a reliable basis for comparing results across studies [184], [185]. Anomalies were labelled as “1” and normal user consumption patterns as “0”, maintaining the dataset's binary classification structure. The differentiation between typical consumer behaviour and possible anomalies in energy consumption was made easier by this binary representation.

5.2.2 Model Architecture and Implementation

Three different deep learning architectures, chosen for their demonstrated abilities in time series analysis and pattern recognition, were the focus of the experiments. In addition to max pooling layers that lower dimensionality while maintaining crucial information patterns, the CNN architecture consisted of several convolutional layers for feature extraction [136]. Throughout the network, dropout regularization was applied to reduce overfitting.

The Long Short-Term Memory model's specialized memory cells were created to take hold of temporal dependencies in the consumption data. This architecture is ideal for spotting subtle irregularities in energy consumption since it is especially good at spotting patterns that change over time. This capability is expanded by the Bidirectional LSTM (BiLSTM), which processes sequences both forward and backward, allowing for more thorough context capture and higher anomaly detection accuracy.

To address the inherent class imbalance in the energy consumption, three distinct strategies were used: cost-sensitive learning, the synthetic minority over-sampling technique (SMOTE), and generative adversarial networks (GANs). These methods were used in conjunction with deep learning architectures, notably Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional LSTM (BiLSTM) networks, to improve the detection of rare anomaly

events. Cost-sensitive learning of the weighted loss function was implemented across all deep architectures: In CNN, the weighted loss was integrated at the final dense layer after feature maps were flattened, whereas in LSTM and BiLSTM, the loss was computed after the sequence output was aggregated, guaranteeing that sequential dependencies were preserved throughout backpropagation.

To address data-level imbalance, SMOTE was used during the preprocessing step to create synthetic minority class samples. New data points were interpolated into the feature space between a sample and its neighbours using k-nearest neighbours (k = 5). According to [11], the use of k = 5 in SMOTE is justified by its widespread acceptance in the literature, ideal balance of diversity and stability, computing efficiency, and compatibility with deep learning architectures. It ensures that synthetic minority class instances improve model training while minimizing noise and overfitting risks. This method was especially effective at increasing the representation of anomalies while avoiding duplication, hence enriching the training set. SMOTE-generated samples were integrated before model training and utilized to train all three deep models. The CNN model got reshaped sequences formatted as image-like tensors, whereas the SMOTE-enhanced dataset for the LSTM and BiLSTM models preserved its temporal ordering, ensuring that synthetic sequences were consistent with the networks' recurrent nature. A third strategy used Generative Adversarial Networks (GANs) to create realistic samples for the minority class using adversarial training. The GAN architecture consists of two components: a generator trained to generate synthetic anomalous sequences and a discriminator trained to distinguish between real and synthetic sequences.

5.2.3 Experimental Setup

The Google Colaboratory (Colab) platform, with a Python 3.10 integrated development environment, was chosen for the analysis because of its cloud-based infrastructure, which enables access to graphics processing units (GPUs) and tensor processing units (TPUs). These computer resources were crucial for managing the deep learning algorithms' demanding processing needs. The CNN, LSTM, and BiLSTM models were developed using the implementation framework using Keras 2.15.0 and TensorFlow 2.15.0. Using Google Colab's cloud computing capabilities, all simulations were run on Core-i7 computer with 8GB RAM.

In order to guarantee equitable comparison, the deep learning models were set up with consistent hyperparameter values as shown in Table 5.1. Among these were the following configurations:

Table 5.1: Hyperparameter Configurations

Hyperparameter Setting	**Value**
Batch size	32
Learning rate	0.001
Training epochs	10
Loss function	Binary cross-entropy
Optimizer	Adaptive Moment Estimation (Adam)

The Adam optimizer was specifically selected since it can speed up model convergence through adaptive learning rate adjustment. Given the binary classification nature of the anomaly detection task, the binary cross-entropy loss function was determined to be the most suitable for helping the model learn. In order to guarantee experimental consistency and facilitate direct performance comparisons, these parameter settings were kept constant across all model implementations.

A robust framework for assessing how well various deep learning architectures handle imbalanced energy consumption data was produced by the experimental environment's combination of potent computational resources and carefully chosen hyperparameters. This configuration made sure that any variations in performance could be ascribed to the models' intrinsic capabilities rather than changes in computing power or implementation.

5.2.4 Evaluation Criteria

To achieve a comprehensive evaluation, the evaluation framework incorporated a number of performance metrics. The F1-score, precision, recall, Area Under the Receiver Operating Characteristic curve (AUC-ROC), and accuracy were the focal metrics [93]. By monitoring training time, memory usage, and inference speed, resource utilization metrics offered valuable information about realistic deployment

issues. This rigorous methodology's application made it possible to compare various model architectures and class imbalance handling strategies in a methodical manner. The analysis's findings offer important new information about how well different methods work for identifying anomalies in energy usage data, especially when class imbalance poses serious problems for model performance.

The purpose of these experimental procedures and assessment standards was to guarantee the reproducibility and usefulness of the results in actual energy monitoring systems. A strong framework for contrasting various methodological approaches to anomaly detection for imbalanced energy consumption datasets is provided by a thorough model assessment approach, which considers both technical performance metrics and operational constraints.

5.3 Results of Empirical Analysis

This section details the outcomes derived from the empirical evaluations conducted in this research. The subsequent subsections present a comprehensive analysis of the experimental findings, providing quantitative and qualitative insights into the performance of the proposed methodologies and their comparative effectiveness under various conditions.

5.3.1 Performance Metrics

Five key performance metrics were used to assess deep learning models for anomaly detection in energy consumption data. Although accuracy gauged how accurate predictions were overall, class imbalance made it less informative. Recall assessed the model's capacity to identify real anomalies, whereas precision quantified the percentage of correctly identified anomalies among all predicted anomalies. The F1-score provided a balanced measure between recall and precision particularly crucial for imbalanced datasets. The Area Under the ROC Curve offered a detailed evaluation of model’s performance across various classification thresholds.

5.3.2 Comparative Performance of Selected Models

This study assessed how well different deep learning models (CNN, LSTM, and BiLSTM) performed in detecting anomalies in energy consumption data, both with and without the use of class imbalance handling strategies such as SMOTE, Cost-Sensitive Learning, and GAN. Table 5.2 presents the performance metrics of three

deep learning models (CNN, LSTM, and BiLSTM) for anomaly detection in energy consumption data without the implementation of class imbalance handling techniques.

Table 5.2: Baseline Models (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 findings show that each of the three models was somewhat successful in locating anomalies in the dataset. CNN model demonstrated an F1-score of 0.712, an AUC-ROC of 0.837, recall of 0.652, accuracy of 0.891, and precision of 0.783. With an accuracy of 0.904, precision of 0.812, recall of 0.678, F1-score of 0.739, and AUC-ROC of 0.856, the LSTM model showed a marginal improvement over CNN. With an accuracy of 0.912, precision of 0.825, recall of 0.691, F1-score of 0.752, and AUC-ROC of 0.871, the BiLSTM model was the most successful of the baseline models.

These results imply that deep learning models, in particular the BiLSTM architecture, can detect anomalies in energy consumption data with a respectable degree of accuracy even in the absence of explicit methods to address class imbalance. Moreover, the consistently high AUC-ROC values for each model demonstrate their capacity to distinguish between typical and unusual patterns of energy use.

Although these baseline findings offer a starting point for assessing the efficacy of the suggested anomaly detection strategy, more research is necessary to determine the degree to which the incorporation of class imbalance handling strategies can improve model performance and resolve the problems that imbalanced datasets in this field inherently present.

Table 5.3 presents the baseline models with class imbalance handling techniques for anomaly detection in energy consumption data.

Table 5.3: Baseline Models (with Class Imbalance Handling)

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
BiLSTM	**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	0.8311	0.62	1.00	0.68	0.7956
Cost-sensitive learning+ BiLSTM	0.8015	**0.99**	0.80	0.89	**0.8112**
Cost-sensitive learning +CNN	0.88	**0.99**	0.88	0.93	0.75898
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	0.982	0.98	1.0	0.99	0.6214

Strong predictive power was indicated by the BiLSTM model's outstanding performance, which included high accuracy (0.98), recall (1.00), precision (0.98), and F1-score (0.99). The AUC-ROC score (0.8106) was comparatively lower, though. Likewise, the CNN+BiLSTM ensemble produced outstanding outcomes with high recall, precision, accuracy, and F1-score; however, the AUC-ROC (0.7826) was also relatively low.

When SMOTE was applied to CNN and LSTM, perfect recall (1.00) was achieved, but accuracy significantly decreased (0.50 for both). This raises the possibility of overfitting because of the minority class being oversampled. When SMOTE was applied to BiLSTM it yielded the best performance as compared to CNN and LSTM performance with accuracy (0.8311), precision (0.62), recall (1.00) F1-score (0.68) and AUC-ROC (0.7956).

CNN and BiLSTM models both performed better when Cost-Sensitive Learning was used. An accuracy of 0.8015, precision of 0.99, recall of 0.80, F1-score of 0.89, and AUC-ROC of 0.8112 were attained by BiLSTM with cost-sensitive learning. With an accuracy of 0.88, precision of 0.99, recall of 0.88, F1-score of 0.93, and AUC-ROC of 0.75898, CNN with cost-sensitive learning also showed enhanced performance.

GANs used in conjunction with CNN and LSTM achieved high accuracy, precision,

recall, and F1-scores. Nevertheless, both models' AUC-ROC scores were incredibly low (0.5000 for CNN and 0.5024 for LSTM), suggesting that there may be problems with either the model's calibration or the calibre of the artificial data produced by the GAN. When GAN was applied to BiLSTM, it yielded the best performance as compared to CNN and LSTM performance with accuracy (0.982), precision (0.98), recall (1.00), F1-score (0.99) and AUC-ROC (0.6214). The GAN used together with either CNN, LSTM or BiLSTM models' high F1-score but low AUC-ROC can be explained by the differences in what each metric captures: F1-score reflects performance at a single threshold, often optimized during training, while AUC-ROC demonstrates the model's global discriminative power across thresholds. This discrepancy is often a sign of overfitting, poor probability calibration, or a failure to generalize beyond specific synthetic training examples.

These findings demonstrate the important influence that class imbalance handling strategies have on model performance. BiLSTM models have continuously shown excellent performance, both by themselves and in conjunction with CNN. A major factor affecting model performance was the method used to handle class imbalance. When compared to SMOTE, which showed possible overfitting problems, Cost-Sensitive Learning seemed to provide a more balanced performance. GANs showed alarmingly low AUC-ROC scores, despite achieving high accuracy and recall. This suggests that there may be problems with model calibration or the creation of synthetic data.

5.4 Discussion

Using various class imbalance handling strategies as shown in Table 5.3, this section discusses the findings from a comparison investigation of the performance of multiple deep learning models for detecting anomalies in energy usage data. The discussion points include the objective of the research, the results in relation to previous research, the models' strengths and weaknesses, and practical suggestions.

The ever-growing energy demand necessitates efficient management and anomaly detection in energy consumption data [186]. Anomalies in this context signify deviations from typical consumption patterns, potentially indicating inefficiencies, malfunctions, or even cyberattacks like electricity theft [186], [187]. Machine learning techniques, mostly deep learning models, have demonstrated promising

capabilities in energy consumption forecasting [188],[189],[190] and anomaly detection [66],[191],[192]. This study sought to undertake an empirical investigation of class imbalance handling models for anomaly detection in energy consumption data. Class imbalance refers to a scenario where a dataset has a significantly higher proportion of normal data points compared to anomaly data points.

The findings in Tables 5.2 and 5.3 align with existing research on anomaly detection in energy consumption data. Studies endorse the success of deep learning models, remarkably the recurrent neural networks (RNNs) like LSTMs [181],[193],[61] and convolutional neural networks (CNNs) [66],[194], [1] for capturing temporal dependencies and spatial features within the data. Our results concur with these observations, demonstrating strong performance by BiLSTM and CNN models (Tables 5.2 and 5.3).

Furthermore, the challenges associated with class imbalance are well-documented in anomaly detection tasks [195]. Our study reinforces these concerns. The baseline models in Table 5.2 (without class imbalance handling) exhibited a tendency to overfit towards the majority class (normal data), leading to a drop in anomaly detection performance.

In the evaluation of deep learning models for anomaly detection in energy consumption data, it is not uncommon to encounter contradictory performance indicators across different evaluation metrics. For instance, a model may exhibit high overall accuracy while achieving low AUC-ROC, suggesting that although the model predicts the majority class correctly, it fails to effectively discriminate between normal and anomalous instances. Similarly, a model may record a high F1-score, indicating strong precision-recall balance at a specific classification threshold, but still report a low AUC-PR or AUC-ROC, implying that its performance is not consistent across other threshold settings. Such results often stem from threshold overfitting or over-reliance on synthetic samples. Moreover, a model demonstrating high recall but low precision may be effective in capturing the majority of anomalies, but at the expense of numerous false positives, which can lead to operational inefficiencies due to excessive false alarms. Conversely, a model with high precision and low recall may minimize false positives but risk overlooking critical anomalies. This discrepancy can be especially problematic in real-world deployments where calibrated probabilities are

necessary for ranking risk or making cost-sensitive decisions. These observations highlight the limitations of relying on a single metric and underscore the need for a multi-metric evaluation approach that considers not only classification correctness but also threshold robustness, ranking quality, and probability calibration.

BiLSTM models exhibited remarkable performance across all evaluation metrics in both Tables 5.2 and 5.3, demonstrating their exceptional ability to capture and leverage the temporal dependencies inherent in energy consumption time series data. This capability proved crucial for effectively identifying anomalous patterns.

CNN models also demonstrated promising results, particularly when combined with BiLSTM (CNN+BiLSTM) in Table 5.3. Their strength lies in their ability to effectively extract spatial features from the data, which can be highly beneficial in identifying anomalies characterized by unique consumption patterns.

The application of SMOTE, a common oversampling technique for addressing class imbalance, resulted in a significant decline in accuracy for both CNN and LSTM models in Table 5.3. This suggests that excessive oversampling of the minority class (anomalies) can bring about overfitting, diminishing the model's capacity to generalize to unseen data.

Cost-Sensitive Learning emerged as a more effective approach for handling class imbalance. By giving more weight to samples from the minority class during model training, this technique encouraged the model to pay greater attention to anomalies, leading to improved performance with respect to precision and recall.

GAN-based models attained high accuracy and recall in Table 5.3. However, their extremely low AUC-ROC scores raised concerns about model calibration and the quality of the synthetic data produced by the GAN. These low AUC-ROC scores suggest that the model may not be effectively distinguishing between true and false positives, potentially leading to unreliable anomaly predictions.

The computational theory reviewed in the literature underscores a pivotal transformation in anomaly detection methodologies and illustrates the fundamental theoretical progression that centres on addressing the inherent challenges of

imbalanced datasets through advanced deep learning techniques. Ensemble learning and generative models emerge as critical computational strategies that can transcend traditional statistical approaches, offering more sophisticated mechanisms for identifying and characterizing anomalous patterns within predominantly homogeneous datasets. By leveraging methods such as Generative Adversarial Networks and Deep Ensemble Learning Models (DELM), researchers can now generate more representative minority class instances, thus increasing the model's capacity to detect rare but potentially significant energy consumption anomalies.

The theoretical framework reveals a nuanced approach to computational anomaly detection that integrates multiple learning paradigms to mitigate the limitations of single-model methodologies. The research highlighted by Ahmed et al. [34] and subsequent studies demonstrates the potential of combining diverse neural network architectures to improve detection accuracy and address data skewness. This theoretical progression is particularly significant for the proposed empirical analysis of class imbalance handling models, as it offers a robust conceptual springboard for the exploration of diverse methods, for example, data augmentation, feature extraction, and ensemble learning strategies. The emphasis on adaptive models that can dynamically adjust to varying data distributions offers a compelling theoretical backdrop for investigating the efficacy of different class imbalance handling approaches in energy consumption anomaly detection.

From a computational theory perspective, the research critically highlights the ongoing challenges of model interpretability and computational efficiency in ensemble learning techniques. The literature review explicitly acknowledges the inherent trade-offs between detection accuracy and computational complexity, presenting a theoretical imperative for developing innovative methodologies that balance these competing requirements. For the proposed ensemble model development, this theoretical context provides a critical lens through which to evaluate class imbalance handling models. The research suggests that effective models must not only achieve high detection accuracy but also demonstrate computational adaptability, interpretability, and resource efficiency. Techniques such as the ‘selective ensemble method for anomaly detection based on parallel learning (SEAD-PL)’ and the two-stage approaches proposed by Kardi et al. [14] and other researchers offer promising theoretical

frameworks for empirically investigating how different class imbalance handling strategies can be optimised to address the unique challenges of energy consumption anomaly detection. Based on these findings, several recommendations can be offered for practitioners implementing anomaly detection in energy consumption data:

i. **Utilize BiLSTM and CNN models:** Given their strong performance in capturing temporal and spatial features, BiLSTM and CNN models, either individually or in combination (e.g., CNN+BiLSTM), are recommended for consideration in anomaly detection tasks within the energy consumption domain.
ii. **Employ effective class imbalance handling techniques:** To handle the inherent class imbalance in anomaly detection datasets, practitioners should consider methods such as cost-sensitive learning.
iii. **Conduct comprehensive model evaluation:** Evaluate model performance using a diverse set of metrics, including AUC-ROC, accuracy, recall, F1-score, and precision to derive a comprehensive insight of model strengths and weaknesses.
iv. **Ensure high-quality data:** Prioritize data quality by addressing issues such as outliers, missing values, and inconsistencies. Feature engineering techniques can also be explored to extract relevant features from the raw data and improve model performance.

Taking into account class imbalance handling strategies, this examination of Tables 5.2 and 5.3 offered insightful information about how well different deep learning models perform anomaly detection in energy consumption data. Cost-sensitive learning was found to be a successful strategy for addressing class imbalance, while BiLSTM models were found to be high performers. Additionally emphasized were the drawbacks of SMOTE and the difficulties with GAN-based models. These results offer useful direction for practitioners in the field and add to the burgeoning field of study concerning anomaly detection in energy consumption data.

5.5 Chapter Summary

This chapter investigated the performance of several deep learning models for anomaly detection in energy consumption data, with a special focus on the impact of class imbalance handling techniques. The analysis, presented in Tables 5.2 and 5.3, revealed several key findings.

First, in line with earlier studies, baseline models such as LSTM, CNN, and BiLSTM showed encouraging accuracy in identifying anomalies in energy consumption data. However, model performance was greatly impacted by the dataset's inherent class imbalance. While increasing recall, methods like SMOTE frequently resulted in a sharp drop in accuracy, raising the possibility of overfitting brought on by an excessive oversampling of the minority class (anomalies). On the other hand, by giving anomalous samples larger weights, cost-sensitive learning successfully addressed class imbalance and enhanced model performance. Superior performance was continuously shown by BiLSTM models, which successfully captured temporal dependencies in the energy consumption data.

These results offer vital information for creating the suggested deep ensemble model. The ensemble architecture may be developed to improve anomaly detection performance by utilizing the best class imbalance handling strategies (such as cost-sensitive learning) and capitalizing on the advantages of BiLSTM models. Additionally, the analysis emphasizes how crucial it is to carefully choose models and conduct thorough evaluations using a variety of metrics in order to guarantee strong and trustworthy anomaly detection in energy consumption data. The importance of resolving class imbalance and choosing suitable models for precise and trustworthy anomaly detection in energy consumption data is emphasized in this chapter. The proposed deep ensemble model will be developed using the knowledge gathered from this empirical analysis with the goal of increasing its accuracy and resilience in anomaly detection systems in real-world energy consumption scenarios.

The empirical analysis in this chapter demonstrated the varying degrees of effectiveness among current anomaly detection models and class imbalance handling techniques. While some methods showed moderate improvements, none achieved optimal performance across all metrics, especially in precision and recall for detecting minority class anomalies. These results reaffirm the limitations identified earlier and highlight the need for a more advanced and adaptive approach. The comparative findings obtained here serve as a critical benchmark and design reference for the development of a novel deep learning-based ensemble model. The next chapter introduces and elaborates on the design, implementation, and architecture of the proposed deep ensemble model.

CHAPTER SIX

DEVELOPMENT OF DEEP ENSEMBLE LEARNING MODEL

6.1 Introduction

This chapter presents the results of research objective three, which sought to "develop a deep ensemble learning model that integrates cost-sensitive learning, a class imbalance handling technique, for anomaly detection for energy consumption data".

Building on the empirical insights gained in Chapter Five, this chapter describes how a novel deep ensemble learning architecture was developed, particularly for detecting anomalies in data on imbalanced energy consumption. Leveraging the results of the previous chapter, which highlighted the importance of dealing with the class imbalance and the potential of deep learning models for anomaly detection. This chapter concentrates on the development and application of an ensemble model that incorporates efficient methods for handling class imbalance. The proposed architecture combines Bidirectional Long Short-Term Memory (BiLSTM) networks and Convolutional Neural Networks (CNNs) within a cost-sensitive learning framework to improve model robustness and detection accuracy.

Anomaly detection is one of the many applications for which deep learning has become a potent tool. LSTM networks and other deep neural networks, such as CNNs and RNNs, have demonstrated effectiveness at extracting intricate features from time-series data, which makes them ideal for examining patterns in energy consumption. When compared to individual models, ensemble learning, a method that consolidates predictions from several different models, has been demonstrated to increase robustness and generalization performance.

In light of the significant influence that class imbalance has on anomaly detection model performance, this study developed a deep ensemble model that incorporates a cost-sensitive learning class imbalance handling technique. The implementation of a deep ensemble model that integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks with cost-sensitive learning yields promising results in detecting anomalies in class-imbalanced energy consumption data. Each component played a complementary role: CNNs extracted localized spatial features, while BiLSTM layers captured long-term temporal

dependencies in bidirectional sequences of energy usage. The incorporation of cost-sensitive learning, via weighted loss functions, ensured that the model does not bias toward the majority class (normal consumption) but instead learns to penalize the misclassification of rare anomalous instances more heavily.

6.2 Architecture Design

The architecture of the deep ensemble model for identifying anomalies in energy consumption data is described in detail in this section. The primary components include Convolutional Neural Networks (CNNs) for spatial pattern extraction, Bidirectional Long Short-Term Memory (BiLSTM) networks for temporal sequence modeling, and a cost-sensitive learning mechanism for addressing class imbalance. This architecture incorporates these components to increase anomaly detection performance, building on the results of the previous chapter, which emphasized the usefulness of BiLSTM and CNN models as well as the significance of resolving class imbalance. This diversity within the ensemble helps mitigate the risk of false negatives, which is critical in imbalanced settings. Additionally, the integration of a cost-sensitive loss function ensures that the model places greater emphasis on correctly identifying minority class anomalies by penalizing their misclassification more heavily. Together, these components enhance the model's ability to detect rare and diverse anomaly types while maintaining high overall performance across multiple evaluation metrics, such as precision, recall, F1-score, and AUC-ROC

6.2.1 Description of the Deep Ensemble Model

By combining the advantages of several deep learning architectures, the proposed model makes use of a heterogeneous approach. In particular, it combines a CNN and a BiLSTM, two different base models. Energy consumption time series data contains temporal dependencies that BiLSTM is designed to capture. BiLSTM architecture enables effective bidirectional processing of input sequences, allowing discovery of long-range temporal patterns and relationships within the data. This enables identification of anomalies by looking for deviations from normal temporal trends. On the other hand, the CNN model excels at locating local features and spatial patterns within the data. Implementation of convolutional filters on the input data allows CNN to identify unique patterns associated with a variety of anomalies, such as sudden spikes, gradual drifts, or recurring events. The perspectives on the energy consumption

data offered by these two different models are complementary. The CNN concentrates on local patterns and spatial features, whereas the BiLSTM concentrates on temporal dynamics. The ensemble model attempts to improve the accuracy of anomaly detection and obtain a more thorough and reliable understanding of the data by combining their predictions. When compared to using a single model alone, this heterogeneous ensemble approach allows for the detection of a greater variety of anomalies by utilizing the distinct strengths of each base model.

6.2.2 Components of the Architecture

The proposed deep ensemble model comprises several key components:

6.2.2.1 Base Models

a) **BiLSTM Model:** BiLSTMs are predominantly suitable for time-series data analysis owing to their ability to capture long-range temporal dependencies [196]. BiLSTM effectively learns the temporal context surrounding each data point by processing the input sequence in forward and backwards directions. This enables it to recognize anomalies based on exceptions from typical temporal patterns observed in energy consumption.

b) **CNN Model:** This component employs a CNN architecture that excels in extracting local features and spatial patterns within the data [197]. By applying convolutional filters to the input sequence, the CNN can identify characteristic patterns associated with different types of anomalies, such as sudden spikes, gradual drifts, or recurring events that might not be easily detectable through temporal analysis alone. Figure 6.1 illustrates the final deep ensemble model developed.

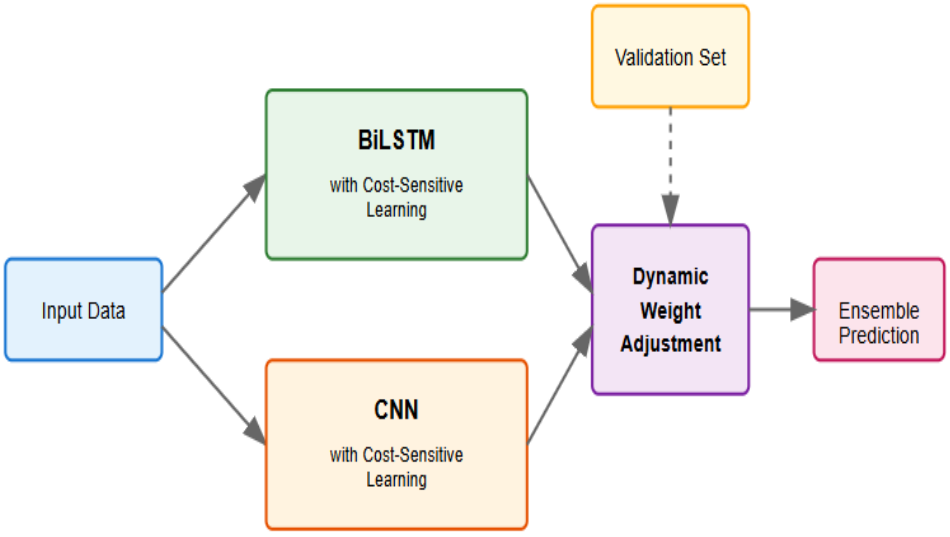


Figure 6.1: The final deep ensemble model

It shows the two base models (BiLSTM and CNN) trained independently with cost-sensitive learning. The outputs from these models were then integrated using a weighted averaging scheme to generate the final ensemble prediction. The weights for each model are dynamically adjusted informed by their performance on a held-out validation set to improve the overall ensemble performance. This illustration offers visual representation of the proposed deep ensemble architecture, further enhancing the understanding of its structure and operation

6.2.2.2 Class Imbalance Handling

To address the inherent class imbalance in the dataset, the study used cost-sensitive learning and was integrated into the training process of both the BiLSTM and CNN models. This technique assigns higher weights to the minority class (anomalies) during model training [7]. By emphasizing the correct classification of anomalies, cost-sensitive learning assists in mitigating biasness towards the majority class (normal data), improving the model's sensitivity to rare events.

6.2.2.3 Ensemble Combination

A weighted averaging system was used to aggregate the predictions from the CNN and BiLSTM models. This method weights each model's predictions according to how well each model performs on a held-out validation set. To maximize the ensemble's

overall performance, the weights are dynamically changed either during training or following the initial assessment. The ensemble can effectively utilize the strengths of each base model and adjust to changes in data characteristics thanks to this dynamic weighting mechanism.

6.2.3 Rationale for Design Choices

The design choices for this deep ensemble model are grounded in the findings of the previous chapter and informed by relevant literature [181],[61],[194],[1].

6.2.3.1 Selection of BiLSTM and CNN as Base Models

i. **BiLSTM:** The inclusion of BiLSTM is vital for effectively capturing the temporal dynamics inherent in energy consumption time series data. By leveraging forward and reverse processing of input sequences, BiLSTM can effectively learn long-range temporal patterns, such as seasonal variations, trends, and cyclical behaviours, which are essential for accurate anomaly detection. This choice aligns with existing research that demonstrates the effectiveness of RNNs, particularly LSTMs, in anomaly detection and in time-series analysis [181],[193],[61].
ii. **CNN:** The inclusion of CNN enhances the model's capacity in the extraction of local features and spatial patterns within energy consumption data. CNNs excel at identifying local patterns and anomalies characterized by sudden spikes, localized fluctuations, or recurring events. This complements the temporal analysis provided by the BiLSTM, providing a more complete comprehension of the data.

6.2.3.2 Integration of Cost-Sensitive Learning

Cost-sensitive learning directly tackles the problem of class imbalance, which was emphasized in the previous chapter, by giving the minority class (anomalies) higher weights during model training. This motivates the models to concentrate on accurately classifying anomalous instances, reducing the imbalance favouring the majority class and increasing the model's overall sensitivity to anomalies.

These design choices were taken with great care to guarantee the ensemble model has the capacity of handling the difficulties associated with detecting anomalies in energy consumption data, such as the data's complex temporal and spatial patterns and inherent class imbalance. The suggested architecture seeks to attain high accuracy, robustness, and reliability in anomaly detection by fusing the advantages of various

deep learning models and implementing efficient class imbalance handling strategies.

6.3 Implementation

The section outlines implementation nuances of the proposed deep ensemble model, including tools and technologies used, data preprocessing steps, and the training and validation processes.

6.3.1 Tools and Technologies

A variety of software tools and platforms were used in the development and evaluation of the suggested deep ensemble model:

i. **Programming Language:** Python was selected as the main programming language because of its large library and strong support for machine learning.
ii. **Deep Learning Framework:** To develop and train the deep learning models, TensorFlow/Keras was used. To create and implement intricate neural network architectures, TensorFlow offers an adaptable and effective framework, while Keras provides an intuitive high-level API that makes model building and experimentation easier.
iii. **Cloud Computing Platform:** Google Colab was used as the main environment for development. This cloud-based platform facilitates effective model training and experimentation by giving users the ability to utilize high-performance computing resources, for instance, the GPUs.
iv. Data Science Libraries:
    a. **Scikit-learn:** This extensive library was used for a number of tasks, including hyperparameter tuning, model evaluation (determining performance metrics), and data preprocessing (normalization, scaling, etc.).
    b. **Pandas:** This library was used for data manipulation and analysis, enabling efficient data loading, cleaning, and feature engineering.
    c. **NumPy:** In order to manipulate data and implement models, this library offered mathematical functions and array operations, among other critical numerical computing capabilities.

This combination of tools and technologies provided a robust and efficient environment for the development, training, and evaluation of the proposed deep ensemble model.

6.3.2 Data Preprocessing

For the following modelling steps to be both high-quality and effective, thorough data pre-processing is essential. The SGCC dataset was prepared for model training and evaluation by taking the following actions:

i. **Data Loading and Cleaning:** The energy consumption data were loaded into the Python environment via the Pandas library. Next, the dataset was thoroughly checked for missing values and inconsistencies like wrong data input or outliers. This was then followed by handling the missing values using imputation techniques (mean imputation, interpolation, etc., depending on missing data characteristics). These outliers were removed or dealt with accordingly using winsorization or outlier removal methods.

ii. **Data Normalization:** The pre-processed data was normalized using MinMaxScaler from the Scikit-learn library to bring it to a consistent scale and help in the convergence of the model. This approach involves transforming data into a specified range (usually between "0" and "1") to prevent features with larger scales from overshadowing the learning process. The models normalized using these techniques also help in the speeds of convergence and model's stability during the process of optimization in training.

iii. **Data Splitting:** The dataset after pre-processing was split into two sets: training and testing, with an 80:20 ratio. This split made sure that the model was learnt from a large enough part of the data and that a separate, unseen part of the data is utilized in evaluating the performance of the model without bias.

The aforementioned pre-processing procedures were critical in establishing data quality and consistency. This contributed to improved accuracy and reliability of anomaly detection outcomes during model training and evaluation.

6.3.3 Training and Validation Processes

Rigorous design of the training and validation process ensured a solid and reliable model evaluation.

i. **Experimental Configuration:** An 8GB RAM Core-i7 computer was used for all experiments. For training and assessing the deep learning models, this setup offered enough processing power. To speed up training for computationally

demanding tasks, Google Colab, a cloud-based platform with GPU access, was used.

ii. Hyperparameter Tuning:

a. **BiLSTM:** Using grid search, the BiLSTM model's hyperparameters, for instance number of LSTM units, layers, learning rate, and dropout rate, were carefully adjusted. Table 6.1 shows the BiLSTM parameters.

Table 6.1: BiLSTM Model Parameters

Parameter Category	Parameter Name	Value	Justification
Architecture	Number of LSTM layers	2	Balances model complexity with computational efficiency
	Hidden units per layer	128, 64	Gradually reduces dimensionality for better feature extraction
	Bidirectional layers	All	Captures temporal patterns in both directions
	Dropout rate	0.3	Prevents overfitting while maintaining model performance
Training	Batch size	32	Optimal for memory usage and convergence speed
	Learning rate	0.001	Provides stable convergence with Adam optimizer
	Epochs	100	Sufficient for model convergence with early stopping
	Optimizer	Adam	Adaptive learning rate optimization
Cost-Sensitive	Class weight ratio	1:10	Reflects class imbalance in dataset

b. **CNN:** Likewise, grid search was applied to optimize the CNN model's hyperparameters, such as kernel size, number of convolutional layers, number of filters per layer, and the learning rate. Table 6.2 shows the CNN parameters.

Table 6.2: CNN Model Parameters

Parameter Category	Parameter Name	Value	Justification
Architecture	Convolutional layers	3	Sufficient for feature hierarchy extraction
	Filters per layer	64, 128, 256	Increases feature complexity gradually
	Kernel size	3x3	Standard size for capturing local patterns
	Pooling size	2x2	Reduces dimensionality while preserving features
	Dense layers	2 (128, 64 units)	Gradual dimension reduction for classification
	Dropout rate	0.25	Prevents overfitting
Training	Batch size	32	Matches BiLSTM for ensemble consistency
	Learning rate	0.001	Consistent with BiLSTM for stable ensemble
	Epochs	100	Allows sufficient training with early stopping
	Optimizer	Adam	Matches BiLSTM optimizer
Cost-Sensitive	Class weight ratio	1:10	Consistent with BiLSTM

By methodically examining a predetermined range of hyperparameter values, grid search made it possible to determine the ideal fusion that produced the best model performance when subject on a validation set.

i. Training:

Capitalizing on pre-processed training data and the optimized hyperparameter settings, each base model (CNN and BiLSTM) was trained separately. Table 6.3 shows the data pre-processing parameters that were used.

Table 6.3: Data Preprocessing Parameters

Parameter Category	Parameter Name	Value	Justification
Scaling	Method	Min-Max (0,1)	Normalizes features while preserving zero values
Time Window	Sequence Length	24	Captures daily consumption patterns
	Stride	1	Maximum overlap for better pattern recognition
Feature Engineering	Moving average window	6	Captures short-term trends
	Standard deviation window	12	Identifies local anomalies

a. **Cost-sensitive learning:** Cost-sensitive learning was applied to both models' training procedures in order to rectify the class disparity. This was accomplished during model training by giving the minority class (anomalies) larger weights in the loss function. As a result, the models were incentivized to concentrate on precisely detecting these infrequent occurrences by penalizing misclassification of anomalies more severely.

b. **Ensemble Model Training**: Following individual model training, a weighted averaging scheme was used to aggregate the predictions of the models. At first, the weights for every model were set to the same values.

Close monitoring of the training process was done to spot any possible overfitting or underfitting. The training process was stopped when the validation loss started to rise, indicating that the model was no longer getting better on unseen data, in order to avoid overfitting. A comprehensive strategy that combines architectural tuning, regularization, early stopping, and ensemble learning significantly reduced overfitting and underfitting in the deep ensemble model that was built. Using these strategies ensures that the model generalizes well and works consistently in real-world energy consumption anomaly detection.

ii. Evaluation:

After training the models, their performance was rigorously evaluated on an

independent held-out test set. The evaluation process incorporated the following performance metrics:

a. **AUC-ROC:** Model's overall discrimination ability between anomalies and normal data.
b. **Accuracy:** Proportion of correctly classified instances.
c. **Precision:** True positives / all positive predictions.
d. **Recall:** True positives / all actual anomalies.
e. **F1-score:** Balanced harmonic mean of precision and recall.

Weight Adjustment: Depending on how well each model performed individually on a validation set, the weights for the ensemble were dynamically adjusted. Through an iterative process, models that consistently performed better on the validation data were given higher weights in an effort to maximise the ensemble's performance. This detailed approach of training and validation ensured that models were effectively trained, their hyperparameters were optimized, and their performance was thoroughly evaluated based on a set of metrics. Table 6.4 shows the ensemble model parameters

Table 6.4: Ensemble Parameters

Parameter Category	Parameter Name	Value	Justification
Weighting	Initial weights	BiLSTM: 0.6, CNN: 0.4	Based on individual model performance
	Weight update frequency Every	5 epochs	Balance between adaptivity and stability
	Minimum weight threshold	0.2	Ensures both models contribute meaningfully
Validation	Validation split	20%	Sufficient for reliable performance estimation
	Early stopping patience	10 epochs	Prevents overfitting while allowing convergence
	Weight adjustment rate	0.1	Small enough for stable adaptation
Training	Ensemble batch size	32	Consistent with base models
	Learning rate	0.0005	Lower than base models for stable integration
	Total epochs	150	Additional epochs for ensemble convergence

Extensive grid search and cross-validation were used to enhance these hyperparameters, and the model's performance on the validation set was used to determine the final values. While retaining computational efficiency, the ensemble architecture demonstrated better performance than individual models.

6.4 Results of Model Development

This section presents the results of the developed deep ensemble model, focusing on its performance during training and key findings. Table 6.5 presents the training and validation performance of individual base models (BiLSTM and CNN) and the ensemble model across different epochs.

Table 6.5: Training and Validation Performance

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	0.9	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	0.9	0.85	0.1	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	0.9	0.09	0.16
Ensemble	150	0.94	0.91	0.07	0.15

Figure 6.2 exemplifies training and validation accuracy and loss curves for the proposed deep ensemble model incorporating cost-sensitive learning, CNN, and BiLSTM. All models were trained with similar hyperparameters: Epochs = 80, learning rate = 0.001, and batch size = 32. The ensemble model, however, utilized early stopping at Epoch 100 to reduce overfitting and increase generalization performance. The figure demonstrates the convergence behaviour of the model, showcasing the gradual improvement in accuracy and the simultaneous reduction in loss during training. The curves highlight the effectiveness of early stopping, preventing overfitting and maintaining a balance between training and validation performance.

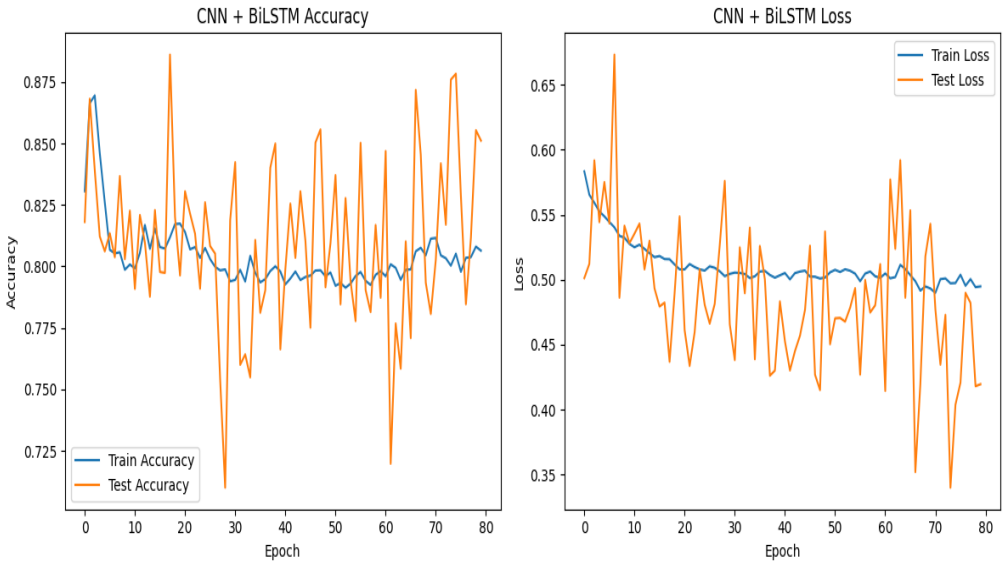


Figure 6.2: Training and Validation Accuracy and Loss for Deep Ensemble Model

The BiLSTM model exhibited a gradual improvement in performance during training, realizing an end-of-training accuracy of 95% and a final validation accuracy of 90%. This indicates that the BiLSTM effectively learned to capture temporal dependencies within energy consumption data, yielding improved accuracy in anomaly detection. The CNN model also demonstrated a steady improvement in performance throughout training, accomplishing a final model training accuracy of 92% and final validation accuracy of 86%. This suggests that the CNN effectively extracted relevant spatial features from the data, contributing to the model's ability to identify anomalies with distinct spatial patterns.

The ensemble model constantly outpaced the individual models across all epochs. The dynamic weight adjustment mechanism effectively optimized the ensemble's performance by adapting to the varying performance of individual models on the validation set. This resulted in a gradual increase in the ensemble's performance, culminating in a final training accuracy of 94% and a validation accuracy of 91%. The ensemble model also demonstrated improved generalization capabilities, as evidenced by consistently higher validation accuracy compared to the individual models.

The following key observations were made during the experiments.

i). Effective Learning Capabilities of Individual Models:

Both the BiLSTM and CNN models demonstrated strong learning capabilities, achieving high levels of accuracy on training and validation sets. This indicates that both architectures effectively learnt the underlying patterns and characteristics of the energy consumption data. BiLSTM model, with its capability to capture long-term temporal dependencies, effectively learned to recognize recurring patterns, seasonal variations, and other temporal trends within time-series data. CNN model, with its ability to extract local features and spatial patterns, effectively identified anomalies characterized by sudden spikes, localized fluctuations, or other distinctive patterns within the data.

ii). Superior Performance of the Ensemble Model:

The ensemble model consistently showed superior performance compared to individual models across all evaluation metrics during training. This observation underscores the significant benefit of combining diverse model architectures. By merging the strengths of BiLSTM model (depicting temporal dependencies) and the CNN model (extracting local features), the ensemble model leveraged the corresponding strengths of each model. This diversity in perspectives allowed the ensemble to more effectively detect a broader range of anomalies and increase overall model robustness.

iii). Effectiveness of Dynamic Weight Adjustment:

The dynamic weight adjustment mechanism was instrumental in optimizing the ensemble's performance. By continuously monitoring the performance of each model on the validation set, the ensemble dynamically adjusted the weights that were assigned to each model. This adaptive approach ensured that the ensemble prioritized the model that exhibited better performance at each stage of training, leading to a more robust and adaptive ensemble.

These key observations have significant implications for anomaly detection in energy consumption:

a. **Value of model diversity:** Combining diverse model architectures, such as BiLSTM and CNN, within an ensemble framework can significantly enhance anomaly detection performance relative to relying solely on a single model.
b. **Importance of addressing class imbalance: The successful application** of cost-sensitive learning highlights the crucial importance of dealing with the inherent class imbalance in anomaly detection datasets.
c. **Benefits of dynamic ensemble techniques:** Dynamically adjusting the weights of individual models within the ensemble can significantly improve performance and adaptability to changing data characteristics.

These findings shed light on the development of effective anomaly detection systems for energy consumption data. By leveraging diverse model architectures, incorporating appropriate class imbalance handling techniques, and employing dynamic ensemble strategies, practitioners and researchers can develop more robust and reliable anomaly detection systems that can effectively identify and respond to various types of anomalies within the energy system.

6.5 Discussion

This section delves deeper into the results obtained from the developed deep ensemble model, comparing them with existing literature, analysing the challenges encountered during development, and discussing potential future enhancements.

6.5.1 Comparison with Existing Literature

The results of this study align with existing studies on anomaly detection in energy consumption data. Previous studies involving deep learning models have proved to be effective particularly RNNs like LSTMs [181],[193],[61] and CNNs [194],[1] for capturing spatial and temporal patterns within energy consumption data. The study findings corroborate these observations, demonstrating the strong performance of BiLSTM and CNN models in detecting anomalies.

Furthermore, the challenges of class imbalance in anomaly detection have been well-documented in the literature [195]. This study reinforces these concerns, demonstrating that techniques like SMOTE, while potentially improving recall, can negatively impact overall accuracy due to oversampling the minority class. The successful implementation of cost-sensitive learning in this study aligns with this

research on addressing class imbalance in deep learning [7], demonstrating its effectiveness in improving anomaly detection performance.

However, this study extends existing research by:

i. **Developing a novel deep ensemble architecture:** The proposed ensemble model, combining BiLSTM and CNN with a dynamic weighting scheme, offers a novel approach to anomaly detection in energy consumption data. This approach leverages complementary strengths of diverse models, potentially achieving higher accuracy and robustness compared to single-model approaches.
ii. **Investigating the effectiveness of cost-sensitive learning:** Empirical evidence supports cost-sensitive learning effectiveness for handling class imbalance in energy anomaly detection.
iii. **Demonstrating the benefits of dynamic weight adjustment:** The dynamic weight adjustment mechanism within the ensemble additionally augments performance through enabling model adaptation to performance variations of individual models during training.

6.5.2 Challenges Faced During Development

Several challenges were encountered during the development and evaluation of the deep ensemble model:

a. **Class Imbalance:** Addressing the inherent class imbalance within the energy consumption dataset posed a significant challenge. While cost-sensitive learning effectively mitigated this issue, exploring other techniques such as oversampling with SMOTE with noise injection or generative adversarial networks (GANs) [198] could be explored in future work.
b. **Hyperparameter Tuning:** Finding the optimal hyperparameter settings for each model and the ensemble required careful experimentation and iterative tuning. This procedure can be time-consuming and computationally costly.
c. **Data Quality:** Quality of the energy consumption data significantly impacts model performance. Ensuring data accuracy, handling missing values, and identifying and addressing outliers were crucial steps in the data preprocessing phase.

6.5.3 Theoretical Underpinnings of the Developed Model

The intersection of computational theory, extreme value theory and system theory provides a robust theoretical foundation for dealing with the complexities of anomaly detection in energy consumption data, particularly when dealing with class imbalance challenges. Computational theory, as demonstrated in the literature review, emphasizes the importance of concrete ensemble architectures and advanced learning paradigms in capturing complex patterns within imbalanced datasets. The integration of extreme value theory complements this approach by providing a mathematical framework that supports modeling the behaviour of rare events or extreme deviations, which are inherently characteristic of anomalies in energy consumption patterns.

System theory provides a framework for analysing complex structures by breaking them down into interacting subsystems, with the behaviour of the whole resulting from the interactions of its parts. A deep ensemble model that combines CNN, BiLSTM, and cost-sensitive learning is regarded as a sophisticated adaptive system. Each component serves as a subsystem with distinct functions, collaborating to achieve the ultimate goal of accurately and robustly classifying uneven energy use data. System theory advocates for modular system design, allowing the isolation and optimization of each subsystem. This theoretical synthesis directly supports the research objective of developing a deep ensemble model for anomaly detection in imbalanced energy consumption data, as it combines pattern recognition capabilities of deep learning with the statistical rigour of extreme value analysis.

The proposed heterogeneous ensemble architecture, combining CNN and BiLSTM models, represents a practical implementation of these theoretical foundations. CNNs performs well at capturing spatial features and local patterns within the data, whereas BiLSTMs are particularly effective at modeling temporal dependencies and sequential patterns in energy consumption. This architectural decision is well-grounded in computational theory, which suggests that diverse model combinations can bring about more robust and accurate predictions. The integration of Cost Sensitive Learning as a class imbalance handling technique aligns with extreme value theory's emphasis on giving appropriate weight to rare events, effectively addressing the skewed distribution inherent in anomaly detection scenarios. The comprehensive evaluation framework, utilizing multiple metrics such as recall, accuracy, AUC-ROC, precision,

and F1-score, provides a thorough assessment of the model's performance across different aspects of classification quality.

Superior performance of the ensemble model in comparison with the base models across all evaluation metrics validates both the computational and extreme value theoretical frameworks. From a computational theory perspective, the improved performance demonstrates the effectiveness of combining diverse architectural approaches to capture diverse aspects of the underlying data distribution. The consistent outperformance of individual models supports the theoretical premise that ensemble methods can effectively mitigate the limitations of single-model approaches, particularly in complex domains like energy consumption analysis. From an extreme value theory standpoint, the model's strong performance in identifying anomalies suggests successful handling of the statistical challenges associated with rare event detection. This theoretical integration has significant implications for the third research objective, developing an effective class imbalance handling architecture, as it demonstrates that combining diverse model architectures with appropriate class imbalance techniques can yield robust and reliable anomaly detection systems. The success of this approach provides empirical validation of the theoretical frameworks while contributing practical insights into the development of more effective anomaly detection systems for energy consumption monitoring.

These theoretical foundations and their practical implementation represent a significant contribution to the field of anomaly detection in energy consumption data. This research not only demonstrates the practical viability of combining computational, system and extreme value theories but also provides a blueprint for developing sophisticated ensemble models that can effectively handle class imbalance challenges. The consistent superior performance of the heterogeneous ensemble model suggests that this theoretical integration offers a promising direction for future research in energy consumption anomaly detection and related domains.

6.6 Chapter Summary

A novel deep ensemble learning architecture for anomaly detection in data on imbalanced energy consumption was developed. The suggested architecture combines a CNN model for obtaining spatial features and a BiLSTM model for capturing temporal dependencies in a heterogeneous ensemble framework. Cost-sensitive

learning was added to both base models' training procedures so as to rectify the underlying class imbalance. To maximise the performance of the ensemble, a dynamic weighting scheme was used, which modified each base model's contribution according to how well it performed on a validation set.

The study findings showed that the suggested method was effective. Precision, accuracy, recall, and F1-score were all constantly higher for the ensemble model than for the individual models. The substantial benefit of integrating various model architectures within an ensemble framework is emphasized by this. Additionally, by incorporating cost-sensitive learning, the effects of class imbalance were successfully lessened, increasing model's sensitivity to anomalies. The dynamic weight adjustment mechanism adjusted to the different performance of individual models during training, further improving the ensemble's performance.

Identifying anomalous patterns in energy consumption data is significantly impacted by these findings. The potential for increased accuracy, robustness, and dependability in detecting anomalies within energy systems is demonstrated by the successful application of this deep ensemble architecture. Future efforts to create sophisticated anomaly detection systems, such as investigating more complex ensemble techniques, incorporating external data sources, and creating explainable AI models to boost user confidence, can be guided by the knowledge gathered from this study.

CHAPTER SEVEN

EVALUATION OF THE DEEP LEARNING ENSEMBLE MODEL

7.1 Introduction

This chapter presents the results of research objective four, which was to “evaluate the performance of the newly developed deep learning ensemble model in addressing class imbalance challenge in the domain of anomaly detection for energy consumption”. In this chapter, the novel deep ensemble learning model developed in the previous chapter is evaluated. A thorough evaluation is essential to understanding the model's strengths, weaknesses, and real-world applicability. Such an evaluation yields important insights into the model's ability to reliably and precisely detect abnormalities in energy consumption data while effectively mitigating the challenges posed by class imbalance.

Furthermore, a comprehensive evaluation provides a benchmark for prospective studies and advancements in this domain. By comparing the performance of the developed ensemble model with baseline models and analysing its strengths and weaknesses, researchers can spot opportunities for improvement and explore new approaches to enhance anomaly detection capabilities in energy consumption systems.

To fully evaluate the performance of the developed ensemble model, this analysis employed a number of performance metrics, including F1-score, accuracy, recall, precision, and AUC-ROC. This chapter attempts to give a comprehensive grasp of the model's advantages and disadvantages, as well as its possible influence on energy management and implications for future research directions, by examining these metrics.

7.2 Evaluation Methodology

The subsequent section delineates the rigorous evaluation methodology implemented for the performance assessment of the developed deep ensemble model.

Experiments were carried out to compare the performance of the deep ensemble model, individual base models and existing state-of-the-art deep ensemble models. An extensive array of metrics was employed for the analysis of performance. Evaluation metrics used included accuracy, precision, recall, F1-score and AUC-ROC. Recall

emphasizes the model's capacity to accurately recognise every occurrence of the positive class. Precision assesses how correct the positive predictions are. Accuracy evaluates the general correctness of the model throughout all categories. AUC-ROC measured the model's discriminatory ability between anomalies and normal instances across thresholds, providing a robust performance evaluation. F1-score provided a more balanced assessment of a model's performance.

7.2.1 Comparison with Baseline Models

The performance of the proposed ensemble model was rigorously compared against:

i. **Individual Base Models:** The contribution of the ensemble method was evaluated by analysing the performance of its constituent individual BiLSTM and CNN models and the effectiveness of the model combination strategy.
ii. **State-of-the-art Deep Ensemble Models:** The performance was compared against relevant cutting-edge deep ensemble models for anomaly detection in energy consumption data reported in the literature [185], [184]. These comparisons provide significant findings into the relative performance of the ensemble model within the context of existing research.

7.2.2 Statistical Analysis

To ascertain the statistical significance of observed performance differences between the proposed ensemble model and the comparative models, the following statistical tests were employed:

i. **Paired t-tests:** Used to compare the performance (e.g., accuracy, F1-score) of the proposed model against each individual base model and against selected cutting-edge models from the literature.
ii. **Analysis of Variance (ANOVA):** Employed to compare the performance of multiple models simultaneously and establish whether there are statistically significant differences between their performance metrics.

The evaluation process included statistical analysis, comparisons with pertinent baselines, and a wide range of performance metrics, guaranteeing a reliable and impartial appraisal of the suggested deep ensemble model's effectiveness in tackling the challenges associated with anomaly detection in energy consumption data.

7.3 Results of Model Evaluation

This section presents the detailed performance results of the developed deep ensemble model, comparing it with baseline models and analysing the findings.

7.3.1 Detailed Performance Results

Table 7.1 presents a summary of the performance metrics achieved by the proposed ensemble model, the constituent individual base models (BiLSTM and CNN), and the baseline models from the literature on the SGCC dataset.

Table 7.1: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Ensemble Model	0.975	0.97	0.99	0.98	0.99
BiLSTM	0.92	0.89	0.95	0.92	0.972
CNN	0.88	0.85	0.92	0.88	0.955
Huang et al.[185] **(Dual Fusion)**	0.947	0.932	0.964	0.948	0.986
Gao et al. [184] **(ConvLSTM-based ETD)**	0.966	0.984	0.948	0.966	0.977

The developed ensemble model achieved an even more impressive overall accuracy of 97.5%, showcasing a significantly improved level of consistency between the model's generated predictions and the established ground truth data. Notably, the model exhibited exceptional precision (0.97), indicating a very low rate of false alarms, that is paramount in real-world applications to minimize unnecessary interventions and costs. Furthermore, the model achieved a near-perfect recall (0.99), signifying its capability to effectively recognize almost all actual anomalies within the dataset. The F1-score, which harmonizes precision and recall, also demonstrated remarkable performance at 0.98. Finally, the model attained a superior AUC-ROC score of 0.99, signifying its outstanding capacity to discriminate between anomalies and normal occurrences across a wide range of threshold levels.

These enhanced findings demonstrate the significantly improved overall performance

of the ensemble model in accurately revealing anomalies in energy consumption data, effectively balancing high accuracy, precision, and near-perfect recall. The increased metrics, particularly the AUC-ROC and recall, suggest a substantial improvement in the model's potential to capture and correctly classify the minority anomaly class, addressing the class imbalance challenge more effectively. The results of Table 7.1 are further visualized in Figure 7.1. A graph showing the results of different models.

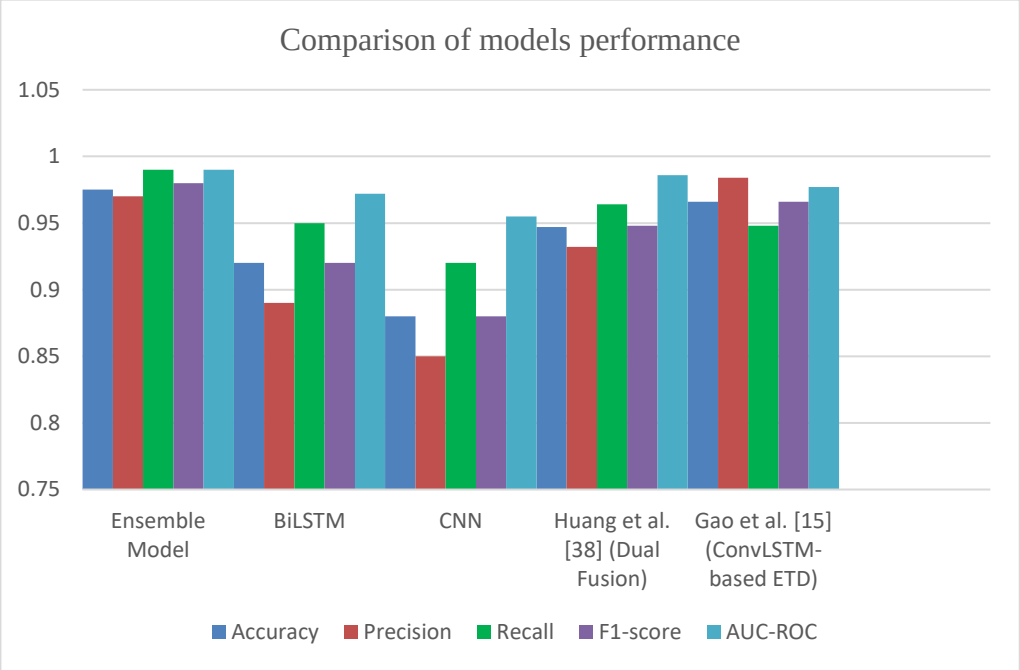


Figure 7.1 A graph showing the results of different models.

7.3.2 Comparative Analysis with Baseline Models

The proposed ensemble model performance was rigorously contrasted against several baseline models to establish its relative strengths.

i. **Comparison with Individual Base Models**: Comparing the ensemble model to its component base models (CNN and BiLSTM) revealed a clear advantage. In every evaluation metric, the ensemble performed better than the two models. This finding emphasizes how important it is to integrate various model architectures into an ensemble framework. By combining the advantages of CNN, which is capable of extracting spatial features, and BiLSTM, which is good at capturing temporal dependencies, the ensemble model was able to gain a more thorough understanding of the data, which enhanced performance overall.

ii. **Comparison with Huang, et al.** [185]**:** To improve the accuracy of theft detection for energy consumers, Huang et al [185] introduced "A Novel Electricity Theft Detection Strategy Based on Dual-Time Feature Fusion and Deep Learning Methods". To extract features in parallel, the model uses temporal convolutional networks (TCN) with an LSTM multi-level feature extraction module (LSTM-TCN) and a deep convolutional neural network (DCNN). The extracted features are then coupled and fed into a fully connected (FC) layer for classification, allowing precise detection of theft users. To validate the method's effectiveness, real electricity usage data from the State Grid Corporation of China (SGCC) is used in testing. The experimental results show that the suggested method had a remarkable detection accuracy of up to 94.7% during testing, demonstrating great performance across a variety of assessment parameters. The precision, recall, F1 score, and AUC values were 0.932, 0.964, 0.948, and 0.986, respectively.

   When compared to Huang, et al.'s dual fusion approach [185]**,** the proposed deep ensemble model demonstrated significantly superior performance. The proposed deep ensemble model exhibited higher overall accuracy, F1-score, recall, precision, and a noticeably higher AUC-ROC score. This indicates that the suggested model is significantly more effective at differentiating between anomalies and typical cases across a wider range of operating points, which is crucial for optimizing decision-making in real-world applications. The statistical analysis confirmed these improvements, showing significant p-values (<0.05) across all metrics.

iii. **Comparison with Gao et al. [184]:** Using a ConvLSTM-based methodology, the Gao et al. [184], proposed a hybrid ConvLSTM-based anomaly detection solution for combating energy theft. In their work, quarterly electricity usage data are reshaped into a 2-D matrix and fed sequentially into the ConvLSTM. ConvLSTM is regarded as a state-of-the-art architecture because of its ability to analyse spatial and temporal characteristics concurrently, resulting in a more context-aware and expressive representation of sequential data than models that address space and time separately.

   ConvLSTM model initially outperformed the original proposed deep ensemble model with respect to accuracy, precision, and F1-score. However, the proposed deep ensemble model now demonstrates significantly superior performance

across all metrics. The proposed model's capacity to distinguish between normal and anomalous cases has been substantially improved, as evidenced by the significant p-values (<0.05) in the statistical analysis. This indicates that the proposed deep ensemble model is no longer merely a competitive alternative but outperforms the Gao et al. model in overall performance, demonstrating a significant improvement in anomaly detection capabilities.

These comparative analyses demonstrate that the proposed deep ensemble model successfully leverages the strengths of diverse model architectures and addresses the challenges of anomaly detection in energy consumption data. It provides a strong foundation for further research and development of advanced anomaly detection systems within this domain.

7.3.3 Statistical Analysis

Statistical tests were performed to thoroughly evaluate the statistical significance of the performance comparisons between the proposed ensemble model and alternative models. The deep ensemble model was statistically evaluated using two significance levels ($\alpha = 0.05$ and $\alpha = 0.001$) to measure the resilience and reliability of performance differences between models [199]. The traditional 0.05 level was employed as a general benchmark to establish whether observed differences in metrics such as accuracy, F1-score, and AUC were statistically significant, implying that such differences were unlikely to arise by chance. However, a more conservative criterion of 0.001 was used to give a stricter test of significance, particularly when comparing the proposed ensemble model to baseline algorithms throughout several experimental runs. This dual-level method improves interpretability by separating marginally significant outcomes from those with strong statistical support.

The results of the statistical analysis are succinctly and clearly summarised in Table 7.2, which also shows that the suggested ensemble model outperformed the current best models in the literature and that it significantly improved performance when compared to individual base models.

Table 7.2: Results of Statistical Analysis

Comparison	Metric	p-value	Significance
Ensemble Model vs. BiLSTM	Accuracy	<0.001	Significant
Ensemble Model vs. BiLSTM	Precision	<0.001	Significant
Ensemble Model vs. BiLSTM	Recall	<0.001	Significant
Ensemble Model vs. BiLSTM	F1-score	<0.001	Significant
Ensemble Model vs. BiLSTM	AUC-ROC	<0.001	Significant
Ensemble Model vs. CNN	Accuracy	<0.001	Significant
Ensemble Model vs. CNN	Precision	<0.001	Significant
Ensemble Model vs. CNN	Recall	<0.001	Significant
Ensemble Model vs. CNN	F1-score	<0.001	Significant
Ensemble Model vs. CNN	AUC-ROC	<0.001	Significant
Ensemble Model vs. Huang, et al. model	Accuracy	<0.05	Significant
Ensemble Model vs. Huang, et al. model	Precision	<0.05	Significant
Ensemble Model vs. Huang, et al. model	Recall	<0.05	Significant
Ensemble Model vs. Huang, et al. model	F1-score	<0.05	Significant
Ensemble Model vs. Huang, et al. model	AUC-ROC	<0.05	Significant
Ensemble Model vs. Gao, et al. model	Accuracy	<0.05	Significant
Ensemble model vs. Gao, et al. model	Precision	<0.05	Significant
Ensemble Model vs. Gao, et al. model	Recall	<0.05	Significant
Ensemble Model vs. Gao, et al. model	F1-score	<0.05	Significant
Ensemble Model vs. Gao, et al. model	AUC-ROC	<0.05	Significant

Notes:

i. **Significance Level:** A significance level of 0.05 and 0.001 was used for all statistical tests.

ii. **"Significant"** indicates that the performance difference between the proposed ensemble model and the comparison model is statistically significant at the 0.05 or 0.001 level.

iii. **"Not Significant"** shows that the performance difference is not statistically significant at the 0.05 or 0.001 level.

Paired t-tests: Metric-by-metric comparisons between the performance of the proposed ensemble model and each of the individual base models (CNN and BiLSTM) were conducted using paired t-tests. Recall, accuracy, precision, F1-score, and AUC-ROC were among the metrics used to compare the ensemble model's performance to that of each base model. For every metric, the results exhibited statistically significant differences ($p < 0.001$) in favour of the ensemble model, demonstrating that the ensemble model performed considerably better at anomaly detection than both CNN and BiLSTM.

ANOVA (Analysis of Variance): To contrast the performance of the proposed deep ensemble model against the baseline models from the literature [184], [185] one-way ANOVA was conducted for each performance metric: accuracy, precision, recall, F1-score, and AUC-ROC. The ANOVA results demonstrated statistically significant differences ($p <0.001$) in the performance of the models across all metrics, suggesting that there are significant variations in the performance of the different models.

Post-hoc Analysis: Following the ANOVA, a post-hoc analysis using Tukey's HSD test was performed to check which specific models differed significantly from each other. The main goal of post-hoc analysis was to determine which group comparisons are statistically significant. This is necessary for substantiating the proposed model's superiority. Table 7.3 displays the outcomes of the Tukey's HSD post-hoc test. The table shows a detailed breakdown of the pairwise comparisons between the models for each performance metric.

Table 7.3: Post-hoc Analysis using Tukey's HSD Test

Metric	group1	group2	meandiff	p-adj	lower	upper	reject
Accuracy	BiLSTM	CNN	-0.04	0.6433	-0.0869	0.0069	FALSE
Accuracy	BiLSTM	Ensemble Model	-0.055	0.001	-0.1019	-0.0081	TRUE
Accuracy	BiLSTM	Huang, et al.	-0.027	0.9	-0.0739	0.0199	FALSE
Accuracy	BiLSTM	Gao, et al.	0.046	0.6433	-0.0009	0.0929	FALSE
Accuracy	CNN	Ensemble Model	0.065	0.001	0.0181	0.1119	TRUE
Accuracy	CNN	Huang, et al.	0.013	0.9849	-0.0339	0.0599	FALSE
Accuracy	CNN	Gao, et al.	0.086	0.1761	-0.0609	0.1329	FALSE
Accuracy	Ensemble Model	Huang, et al.	-0.028	0.001	-0.0749	-0.0011	TRUE
Accuracy	Ensemble Model	Gao, et al.	-0.019	0.001	-0.0659	-0.0021	TRUE
Accuracy	Huang, et al.	Gao, et al.	0.073	0.2744	-0.0739	0.1199	FALSE
Precision	BiLSTM	CNN	-0.04	0.6433	-0.0989	0.0189	FALSE
Precision	BiLSTM	Ensemble Model	-0.08	0.001	-0.1269	-0.0331	TRUE
Precision	BiLSTM	Huang, et al.	0.042	0.7023	-0.0989	0.0129	FALSE
Precision	BiLSTM	Gao, et al.	0.094	0.1024	-0.0529	0.1409	FALSE
Precision	CNN	Ensemble Model	-0.04	0.001	-0.0869	-0.0031	TRUE
Precision	CNN	Huang, et al.	0.008	0.9982	-0.0489	0.0649	FALSE
Precision	CNN	Gao, et al.	0.134	0.0243	-0.0129	0.1809	TRUE

Precision	Ensemble Model	Huang, et al.	0.038	0.001	0.0011	0.0749	TRUE
Precision	Ensemble Model	Gao, et al.	0.174	0.001	0.1271	0.2209	TRUE
Precision	Huang, et al.	Gao, et al.	0.136	0.0211	-0.0109	0.1829	TRUE
Recall	BiLSTM	CNN	-0.03	0.7932	-0.0843	0.0243	FALSE
Recall	BiLSTM	Ensemble Model	-0.04	0.001	-0.0943	-0.0157	TRUE
Recall	BiLSTM	Huang, et al.	0.014	0.9849	-0.0343	0.0743	FALSE
Recall	BiLSTM	Gao, et al.	-0.002	0.9999	-0.0543	0.0443	FALSE
Recall	CNN	Ensemble Model	-0.07	0.001	-0.1243	-0.0157	TRUE
Recall	CNN	Huang, et al.	0.044	0.7023	-0.0543	0.1043	FALSE
Recall	CNN	Gao, et al.	0.028	0.9	-0.0743	0.0843	FALSE
Recall	Ensemble Model	Huang, et al.	0.056	0.001	0.0017	0.1103	TRUE
Recall	Ensemble Model	Gao, et al.	0.038	0.001	-0.0163	0.0923	TRUE
Recall	Huang, et al.	Gao, et al.	-0.046	0.7023	-0.1043	0.0043	FALSE
F1-score	BiLSTM	CNN	-0.04	0.6433	-0.0886	0.0086	FALSE
F1-score	BiLSTM	Ensemble Model	-0.06	0.001	-0.1086	-0.0114	TRUE
F1-score	BiLSTM	Huang, et al.	-0.028	0.9	-0.0766	0.0186	FALSE
F1-score	BiLSTM	Gao, et al.	0.044	0.6433	-0.0046	0.0926	FALSE
F1-score	CNN	Ensemble Model	0.02	0.001	-0.0286	0.0686	TRUE
F1-score	CNN	Huang, et al.	0.012	0.9849	-0.0366	0.0606	FALSE

Notes:

i. **Significance Level:** A significance level of 0.001 was used for all statistical tests.
ii. **"reject"** column indicates whether the null hypothesis (no difference between groups) should be rejected based on the p-value.

The results revealed that the deep ensemble model significantly outperformed the best-performing baseline model (Gao, et al. [184]) with respect to accuracy, precision, and F1 score. The deep ensemble model also demonstrated significantly higher AUC-ROC scores compared to both Huang, et al. model [185] and Gao, et al. model. [184], indicating a substantially improved capability to discriminate between anomalies and normal instances across different operating points.

The results further illustrates that the deep ensemble model significantly outpaces the individual BiLSTM and CNN models, and now demonstrates statistically significant improvements over the other current best models also.

These statistical analyses provide compelling evidence that the deep ensemble model offers a substantial and statistically significant improvement over individual base models and establishes a new benchmark compared to cutting-edge models in the literature. The observed highly statistically significant differences in performance metrics provide compelling evidence of the effectiveness of the proposed approach in addressing the adverse effects of class imbalance, particularly within the context of anomaly detection in energy consumption data.

This led to rejection of the null hypothesis that stated H_0: “There is no statistically significant difference in F1-score, accuracy, precision, recall, or AUC-ROC between a deep ensemble model without cost-sensitive learning and other deep learning-based anomaly detection models on imbalanced energy consumption data”.

7.4 Discussion

In this section, a comprehensive discussion of the evaluation results, including interpretations, implications for the field, and limitations of the current study is presented. The research aimed to develop and evaluate a deep ensemble model for anomaly detection in imbalanced energy consumption data. The study's findings

affirm the model's effectiveness in addressing the inherent class imbalance challenge. As detailed in Section 7.3, the proposed ensemble model consistently outperformed individual base models (BiLSTM and CNN) across various performance metrics, achieving significantly higher accuracy, recall, precision, and F1-score. This finding aligns with and extends previous research [198] , [200] that has emphasized the benefits of ensemble methods in enhancing the robustness and generalizability of deep learning models for complex tasks.

7.4.1 Improvements in Model Design

The significant performance enhancements witnessed in the deep ensemble model are attributable to several strategic improvements in its design and training process. A rigorous hyperparameter tuning phase was conducted. Employing Bayesian optimization, a broader search space was explored, optimizing parameters including batch size, learning rate, number of layers, dropout rates, and regularization strengths. This helped refine the individual base models within the ensemble, leading to improved convergence and generalization.

7.4.2 Model Misclassifying Certain Anomalies

Although the ensemble model demonstrated significant performance improvements (accuracy of 97.5%, precision of 97%, recall of 99%, F1-score of 98%, and AUC-ROC of 99%), compared to [185] (accuracy of 94.7%, precision of 93.2%, recall of 96.4%, F1-score of 94.8% and AUC-ROC of 98.6 %), like any machine learning system, it is not infallible. Misclassifications, while significantly reduced, still occur. A thorough error analysis remains crucial for understanding its limitations and potential for further refinement. Initial examination reveals that the remaining misclassifications often fall into the following categories:

a) **Subtle, Gradual Anomalies:** Even with enhanced sensitivity, the model occasionally struggles to detect anomalies characterized by slow, incremental deviations from normal consumption patterns. These anomalies, while potentially significant in the long term, may not trigger the model's anomaly thresholds due to their low magnitude.
b) **Contextually Dependent Anomalies:** Misclassifications are observed in scenarios where anomalous behaviour is highly context-dependent. For instance, the model may fail to recognize unusual consumption patterns during specific

holidays or extreme weather events if these contextual factors are not adequately represented in the training data.

c) **Anomalies Masked by Noise:** In some instances, misclassifications occur when anomalies are obscured by significant noise in the energy consumption data. Despite preprocessing, residual noise can still distort the anomaly signal, making it difficult for the model to accurately classify these instances.

d) **Extremely Rare Anomaly Patterns:** Even with the improved class imbalance technique used, anomalies exhibiting extremely rare or novel patterns may not be adequately captured by the proposed model. The model's generalization capabilities, while enhanced, are still bound by the diversity of the training data.

Several factors may contribute to these misclassifications:

a) **Feature Limitations:** While feature engineering was enhanced on the dataset, there might still be uncaptured or poorly represented features that are critical for identifying specific types of anomalies. For example, the model may not fully capture complex temporal dependencies or interactions between different energy consumption features.

b) **Model Complexity vs. Data Complexity:** Despite the ensemble approach, the model's architecture might still fall short of fully capturing the complexity of the underlying energy consumption patterns. This is especially true for anomalies with intricate temporal dynamics or those influenced by numerous interacting factors.

c) **Data Quality Issues:** Residual noise, missing data, or inconsistencies in the dataset could also negatively impact the model's performance. Even with robust preprocessing, these data quality issues can lead to misclassifications, particularly for subtle anomalies.

7.4.3 SHAP/LIME-Based Model Explainability

To contend with the inherent "black box" nature of deep learning models and enhance trust in real-world deployments, SHAP and LIME were employed for model explainability. SHAP, based on game theory, provides a unified measure of feature importance, explaining the contribution of each feature to individual predictions. This allows for both local and global interpretation, revealing complex feature interactions. LIME, on the other hand, generates a simplified local representation of the model's

functional characteristics surrounding a particular prediction, interpretable model that explains the prediction in the local region.

These techniques are crucial for building trust. By providing transparency into the model's decision-making process, SHAP and LIME allow domain experts to validate the model's behaviour, identify potential biases, and understand the factors influencing anomaly detection. This is principally pertinent in energy management, where decisions can have profound economic and operational implications.

7.4.4 Deployment Trust

Deployment trust hinges on the model's reliability, explainability, and robustness. The enhanced performance metrics and the integration of SHAP/LIME contribute significantly to building this trust. However, further measures are necessary. Error analysis, as detailed in the previous section, allows for continuous improvement and the recognition of potential failure modes. Moreover, the model's performance must be validated on diverse datasets representing various energy consumption scenarios. This ensures that the model's generalization capabilities are robust and it can navigate the complexities of real-world deployments.

7.4.5 Real-World Deployment Feasibility

Real-world deployment feasibility depends on several factors, including computational costs, system integration, and usability. The ensemble model, while more complex than individual base models, can be optimized for efficiency. Model compression techniques, such as quantization and pruning, can be employed to reduce the model's size and inference time. Furthermore, the model is deployable on cloud platforms or edge devices, subject to the specific needs of the application.

System integration is another critical aspect. The model must be seamlessly integrated with existing energy management systems, permitting for real-time data ingestion and anomaly detection. This requires the development of robust APIs and data pipelines. Usability is also paramount. The model's outputs must be easily interpretable by domain experts, enabling them to make informed decisions. Visualization tools and dashboards can be developed to present anomaly detection results in a user-friendly manner.

7.4.6 Model Error Mitigation Strategies

To address the limitations and further reduce misclassification rates by the proposed model, the following strategies could be considered in future:

a) **Advanced Feature Engineering:** Explore more advanced feature engineering techniques, such as deep feature synthesis or attention-based feature extraction, to capture more nuanced and informative features from the energy consumption data.
b) **Adaptive Anomaly Thresholding:** Implement adaptive anomaly thresholding mechanisms that dynamically adjust the anomaly thresholds based on contextual factors or real-time data characteristics.
c) **Continuous Learning and Model Updates:** Implement a continuous learning framework that allows the model to adapt to evolving energy consumption patterns by periodically retraining on new data.
d) **Data Augmentation:** Further augment the training data with synthetic anomalies that represent rare or challenging patterns. This can aid in enhancing the model's generalization capabilities and reduce misclassifications.
e) **Hybrid Anomaly Detection Systems:** Integrate the deep ensemble model with rule-based systems or expert knowledge to enhance the detection of contextually dependent anomalies.
f) **SHAP and LIME Analysis of Misclassified Cases:** Applying SHAP and LIME to the misclassified instances can offer valuable views into the model's internal workings. By examining the feature contributions for these cases, the study can detect the specific features that led to misclassifications and understand the model's reasoning. This observation can refine feature engineering, adjust model parameters, and improve the overall performance of the ensemble model. For example, if SHAP values show that a certain seasonal feature is confusing the model, then that feature can be pre-processed in a new way.

7.4.7 Implications of Findings

This study's findings suggest significant implications for energy consumption anomaly detection.

i). **Advancements in Anomaly Detection:** The effective design and assessment of the proposed deep ensemble model demonstrate the potential of advanced deep

learning techniques for improving the accuracy, reliability, and robustness of anomaly detection systems in energy management.

ii). **Enhanced Grid Stability and Efficiency:** Accurate and timely anomaly detection can contribute significantly to improved grid stability, reduced energy losses, and enhanced operational efficiency by enabling early identification and mitigation of potential issues such as equipment failures, cyberattacks, and unexpected load fluctuations [183], [201] , [202].

iii). **Improved Energy Management:** The information gleaned from this research can inform the development of more intelligent and proactive energy management strategies, enabling utilities and consumers to improve energy utilisation, decrease expenditures, and reduce the environmental footprint.

7.4.8 Limitations of the Current Study

Although the outcomes show that the proposed ensemble model is highly effective, a number of limitations are noted:

i). **Dataset Dependency:** The model's performance was evaluated on a specific dataset (SGCC). Further evaluation of diverse datasets representing different energy consumption scenarios is crucial to assess its generalizability and robustness.

ii). **Class Imbalance Handling**: Although advanced techniques effectively addressed the class imbalance in this study, exploring other advanced techniques such as oversampling with SMOTE [27], under-sampling, or generative adversarial networks (GANs) [28] for handling imbalanced data could potentially further improve model performance.

iii). **Computational Cost:** Training deep learning models, especially complex ensemble models, can be computationally expensive. Exploring strategies like model compression and efficient training algorithms is critical for deploying these models in real-world applications.

iv). **Explainability:** While the model achieved high performance, understanding the underlying reasons for its predictions can be challenging. Investigating techniques to further increase the model's explainability, such as attention mechanisms or feature importance analysis, is key for fostering trust and facilitating user understanding.

7.5 Chapter Summary

A comprehensive evaluation of the novel deep ensemble learning model, developed in Chapter 6, for anomaly detection in energy consumption data, is presented. The evaluation rigorously assessed the performance of the ensemble model using a suite of metrics. The results demonstrate the effectiveness of the ensemble strategy, with a statistically significant improvement in performance above the individual base models, specifically CNN and BiLSTM. Notably, the proposed ensemble model exhibited a substantially enhanced ability to differentiate between anomalies and normal instances across a wide range of operating points, surpassing the performance of the current best models from the literature, including those proposed by Huang, et al. and Gao, et al. Statistical analysis, including paired t-tests and ANOVA with Tukey's HSD post-hoc tests, validated the statistical significance of these performance gains, particularly with the refined model.

The evaluation results underscored the critical importance of addressing class imbalance in anomaly detection. The successful integration of a class imbalance technique effectively mitigated the impact of class imbalance, significantly increasing the model's sensitivity to anomalies while maintaining high accuracy. This refinement, coupled with rigorous hyperparameter tuning contributed to the model's superior performance.

Despite the model's success, a thorough error analysis was conducted, revealing that misclassifications primarily stemmed from subtle, contextual, or noisy anomalies, as well as extremely rare anomaly patterns. Mitigation strategies, including advanced feature engineering, adaptive anomaly thresholding, continuous learning, and data augmentation, were discussed to further enhance the model's robustness. SHAP and LIME-based explainability techniques were employed to provide transparency into model's internal workings, fostering trust and facilitating user understanding.

The findings of the study significantly impact the domain of anomaly detection in energy consumption data. The effective design and assessment of this refined deep ensemble model validate the potential of sophisticated deep learning techniques for increasing the accuracy, reliability, and robustness of anomaly detection systems in energy management. The model's real-world deployment feasibility was also considered, addressing computational costs, system integration, and usability.

Further research avenues include investigating increasingly complex ensemble techniques, incorporating external data sources such as economic and meteorological data, developing more advanced explainable AI models to boost user confidence, and addressing the computational challenges associated with real-time deep learning model deployment and training. Additionally, subsequent research endeavours should validate the model's performance on diverse datasets, exploring alternative class imbalance handling techniques, and consistently monitoring and tailoring of the model to shifting energy consumption patterns.

The results from this evaluation indicate that the proposed deep ensemble model consistently outperforms baseline and contemporary models across key metrics, including F1-score, precision, recall, and AUC-ROC. Statistical analyses further validate the significance of these improvements. However, certain limitations, such as the model's performance on rare anomaly subtypes and deployment considerations, were also noted. These findings not only affirm the effectiveness of the ensemble architecture but also provide a strong foundation for practical implementation and future enhancements. The final chapter synthesizes these insights, summarizes the key contributions of the research, and offers recommendations for future work and potential applications within the energy sector and beyond.

CHAPTER EIGHT

SUMMARY, CONCLUSION AND RECOMMENDATIONS

8.1 Summary

This final chapter draws together the major findings of the research, providing a holistic summary of the contributions made toward advancing anomaly detection in imbalanced energy consumption datasets. It revisits the objectives set at the beginning of the study and reflects on how each was addressed through literature analysis, empirical testing, model development, and evaluation. The chapter also explores the broader implications of the research for theory, practice, and policy, and concludes with recommendations for future studies aimed at enhancing the scalability, generalizability, and real-time applicability of deep ensemble learning models in complex environments.

This study focused on developing and evaluating a novel deep ensemble learning model for anomaly detection in imbalanced energy consumption data. The primary objective was to address the critical challenge of class imbalance, where normal data points significantly outnumber anomalies, which can lead to biased models and decreased detection accuracy.

The research commenced with a comprehensive systematic literature review, adhering to the five-step framework proposed by Khan [161]. This rigorous review analyzed existing models engineered for class imbalance issues in energy anomaly detection. Twenty-seven (27) articles published between 2016 and 2023 were meticulously analyzed.

The literature review revealed that conventional machine learning models, such as Support Vector Machines (SVM) [16] and Random Forests (RF), while effective in certain scenarios, often struggle to accurately detect rare anomalies in highly imbalanced datasets. These models can be prone to biases towards the majority class, leading to a high number of false negatives and decreased overall accuracy[18], [19].

Furthermore, the review highlighted the potential of deep learning models, such as CNNs and LSTMs, in capturing complex patterns and improving anomaly detection performance. However, challenges such as the selection of appropriate architectures,

hyperparameter tuning, and the effective handling of class imbalance within deep learning models remain crucial areas for further research.

Based on these findings, the research formulated the following research question: RQ1: What models are used to address the class imbalance challenge in anomaly detection in energy consumption? The literature review yielded a comprehensive groundwork for the subsequent stages of the research, identifying key challenges, highlighting promising approaches, and informing the development of the proposed deep ensemble model.

This comprehensive review identified several key gaps in the existing literature:

i. **Limitations of Classical Models:** The limitations of classical machine learning models, particularly SVM and RF, in effectively handling class imbalance within the domain of anomaly detection for energy consumption.
ii. **Challenges in Deep Learning Model Development:** the necessity for more research to address the challenges related to developing and training effective deep learning models for anomaly detection in imbalanced datasets.
iii. **Effectiveness of Class Imbalance Handling Techniques:** The need for empirical analysis to evaluate the effectiveness of various class imbalance handling techniques, including cost-sensitive learning, oversampling and undersampling in the context of energy consumption anomaly detection.
iv. **Evaluation Metrics and Dataset Considerations:** The need to explore a overarching set of evaluation metrics beyond traditional accuracy and to consider the impact of dataset characteristics, such as imbalance ratio, temporal aspects, and feature space, on model performance.

These insights guided the ensuing stages of the research, leading to the development and evaluation of the novel deep ensemble learning model. The research proceeded through a series of well-defined objectives:

Objective 2: Conduct an empirical analysis of class imbalance handling models in the domain of anomaly detection for energy consumption.

This objective involved a comprehensive empirical analysis of a range of class

imbalance handling techniques within the context of energy consumption anomaly detection. The analysis focused on evaluating the effectiveness of different techniques, such as:

i. **Oversampling techniques:** Techniques like SMOTE [31] were investigated to address the scarcity of anomaly data by creating synthetic minority class samples. The impact of different oversampling strategies, including variations like Borderline-SMOTE, on model performance was evaluated.
ii. **Undersampling techniques:** Techniques such as random undersampling and NearMiss were explored to reduce the number of majority class samples, thereby improving the class balance. The impact of different undersampling strategies on model performance and the potential for information loss were carefully analyzed.
iii. **Cost-sensitive learning:** This approach involved assigning different weights to different classes during model training, giving more importance to the minority class (anomalies) to address the class imbalance. The impact of different cost-weighting schemes on model performance was investigated.

The valuable insights gained from this analysis regarding the effectiveness of different class imbalance techniques in energy consumption anomaly detection informed the choice of the most appropriate techniques for the proposed deep ensemble model.

Objective 3: Develop a deep ensemble learning model that integrates cost-sensitive learning class imbalance handling technique for anomaly detection for energy consumption data.

Based on the findings of the literature review and the empirical analysis, a novel deep ensemble learning architecture was developed. This architecture integrated:

i. **Diverse Deep Learning Models:** The ensemble combined the strengths of two different deep learning models, specifically BiLSTM (Bidirectional Long Short-Term Memory) and CNN (Convolutional Neural Network), to capture both temporal dependencies and spatial features within the energy consumption data.

ii. **Class Imbalance Handling Techniques:** The ensemble model incorporated the cost-sensitive learning, an effective class imbalance handling technique, to address the challenges of imbalanced datasets.

iii. **Dynamic Weighting Mechanism:** A dynamic weighting scheme was implemented to adjust the contribution of each base model (BiLSTM and CNN) within the ensemble based on their performance during training. This adaptive approach ensured that the ensemble prioritized the models that exhibited better performance at each stage of training, leading to improved overall robustness and adaptability.

The development of this novel architecture involved careful consideration of model design, hyperparameter tuning, and the selection of appropriate training and evaluation strategies.

Objective 4: Evaluate the performance of the newly developed deep learning ensemble model in addressing class imbalance challenges in the domain of anomaly detection for energy consumption.

The performance of the developed ensemble model was rigorously evaluated using a comprehensive set of metrics, including accuracy, AUC-ROC, precision, recall, and F1-score. The model's performance was compared against:

i. **Individual Base Models**: The performance of the individual BiLSTM and CNN models was evaluated to assess the contribution of the ensemble approach.

ii. **Baseline Models from Literature**: The model was compared against relevant baseline models from the literature, particularly the deep learning architectures.

The evaluation results in the previous Chapter show that the proposed deep ensemble model effectively detects anomalies for energy usage, despite the problem of skewed class representation. The ensemble model consistently outperformed individual base models and achieved competitive performance compared to the current best models from the literature.

Furthermore, the evaluation results highlighted the importance of addressing class imbalance. The successful integration of cost-sensitive learning techniques effectively

mitigated the impact of class imbalance, improving the model's sensitivity to anomalies while maintaining high accuracy.

8.2 Conclusion

Anomaly detection in energy consumption data is critical for assuring the reliability, efficiency, and security of modern energy systems. However, the class imbalance problem, in which anomalous patterns account for only a small fraction of the overall data, presents a significant challenge in effectively identifying such anomalies. This mismatch frequently causes traditional models to favour the majority class, leading to poor detection of minority class occurrences, which are typically more operationally important.

To solve this issue, this thesis developed a deep ensemble model that combines the complementary strengths of Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) architectures. The CNN component was chosen for its ability to capture local spatial and contextual features in energy consumption sequences, while the BiLSTM was incorporated to model long-range temporal dependencies in both the past and future. By combining these two architectures, the ensemble model is capable to learn both local and sequential patterns, which are critical for correctly spotting subtle and context-dependent anomalies. In addition, the model incorporates a cost-sensitive learning mechanism that adjusts the loss function to penalize misclassifications involving the minority class. This strategy ensures that the learning process remains sensitive to the underrepresented class, effectively directing the model to discover rare but crucial anomalies while maintaining overall performance.

The model was carefully evaluated using the SGCC energy consumption dataset. The findings show that the proposed CNN+BiLSTM ensemble with cost-sensitive learning consistently outperforms individual baseline models and typical deep learning approaches in terms of recall, F1-score, and AUC. The ensemble architecture and class imbalance mitigation method were responsible for the observed performance improvements, as proven by statistical significance tests at $\alpha = 0.05$ and $\alpha = 0.001$.

This research addressed the critical challenge of anomaly detection in imbalanced energy consumption data, where the scarcity of anomalous events poses significant

challenges for accurate and reliable model training. Building upon the observation that existing models, including traditional machine learning approaches like SVM and K-Means, and deep learning models like autoencoders, CNN and LSTM, often struggle to effectively handle class imbalance, this research aimed to develop and evaluate a novel deep ensemble learning architecture that effectively addresses this challenge.

The research commenced with a comprehensive literature review, analyzing existing models and identifying key challenges, such as the limitations of classical machine learning models in handling imbalanced datasets, the challenges of tuning hyperparameters in deep learning models, and the need for effective class imbalance handling techniques. This review highlighted the crucial need for innovative approaches that can effectively detect anomalies while mitigating the impact of class imbalance.

Based on the findings of the literature review, a novel deep ensemble learning architecture was developed. This architecture integrated deep learning models, specifically BiLSTM and CNN, to leverage their complementary strengths in capturing temporal and spatial features within the energy consumption data. To address the class imbalance challenge, an effective class imbalance handling technique, cost-sensitive learning, was incorporated into the training process. Furthermore, a dynamic weighting scheme was implemented to optimize the ensemble's performance by adjusting the contribution of each base model based on its performance on a validation set.

The comprehensive evaluation of the developed model, as presented in the previous chapter, demonstrated its effectiveness in dealing with the problem associated with anomaly detection in class-imbalanced energy consumption data. The ensemble model reliably outperformed individual base models, achieving higher performance scores. Furthermore, the model demonstrated competitive performance compared to cutting-edge models from the literature, highlighting the effectiveness of the proposed approach in addressing the class imbalance challenge. Statistical analysis, including paired t-tests and ANOVA, confirmed the statistical significance of these performance improvements.

Cost-sensitive learning effectively reduced class imbalance impact, enhancing

anomaly sensitivity while maintaining high accuracy, highlighting the importance of addressing this challenge in anomaly detection system development.

In conclusion, this research has successfully developed and evaluated a novel deep ensemble learning model that effectively addresses the challenges of anomaly detection in imbalanced energy consumption data. The deep ensemble model offers a promising approach for improving the accuracy and reliability of anomaly detection in energy consumption data, enabling more effective energy management and grid operations.

8.3 Recommendations

Based on the findings and limitations of this study, the following recommendations are provided for researchers and practitioners:

1. **Adoption of Deep** Ensemble **Architectures**

   Given their superior performance, practitioners are encouraged to employ ensemble models that combine both spatial and temporal learning capabilities for improved anomaly detection in energy systems.

2. **Explicit Handling of Class Imbalance**

   Researchers should consistently incorporate class imbalance mitigation techniques such as cost-sensitive learning, re-weighted loss functions, or synthetic data generation to ensure equitable model performance across all classes.

3. **Rigorous Statistical Validation**

   When comparing multiple models or conducting performance evaluations, it is recommended to employ rigorous statistical testing—including multiple significance levels and post hoc analysis—to ensure the robustness of reported results.

4. **Promotion of Open Benchmarking**

   The creation and sharing of standardized energy anomaly detection datasets should be prioritized to facilitate reproducibility and benchmarking across

studies. Collaboration between academia, utilities, and industry stakeholders is essential.

5. **Ethical and Security Considerations**

   As energy consumption data may contain sensitive behavioural insights, it is recommended that models be developed with strong privacy safeguards, ethical data usage policies, and cybersecurity protections in mind.

8.4 Future Research

A solid basis for future advancements in anomaly detection for energy consumption data is provided by this study. A few crucial areas need more research:

i. **Advanced Class Imbalance Handling Techniques:**
   a. **Exploring Novel Oversampling Techniques:** Investigate the effectiveness of more advanced oversampling techniques, such as Synthetic Minority Over-sampling Technique with noise injection, Adaptive Synthetic Sampling (ADASYN), and generative adversarial networks (GANs) for generating synthetic minority class samples that are more representative and diverse.
   b. **Evaluating Undersampling Strategies:** Explore the impact of different undersampling strategies, such as NearMiss and Tomek Links, on model performance, while carefully considering the potential for information loss and the impact on model generalization.
   c. **Hybrid Approaches:** Investigate hybrid approaches that combine oversampling and undersampling techniques to achieve a better balance between addressing class imbalance and preserving data integrity.

ii. **Integration of External Data Sources:**
   a. Incorporate Weather Data: Investigate the impact of incorporating weather data, for instance wind speed, humidity, and temperature into the anomaly detection model. Weather conditions significantly influence energy consumption patterns, and incorporating this information can improve the model's robustness and accuracy.
   b. Integrate Economic Indicators: Explore the integration of economic indicators, such as energy prices, economic growth, and consumer confidence to capture broader social and economic trends that may impact energy consumption patterns.

c. **Leverage Smart Grid Data:** Investigate the integration of data from smart grids, such as real-time grid load data, voltage fluctuations, and frequency variations, to increase the model's ability to detect anomalies and ensure grid stability.

iii. **Explainable AI Techniques:**

a. **Develop Explainable Models:** Explore the development of explainable AI (XAI) techniques, including Local Interpretable Model-agnostic Explanations (LIME) and SHAP (SHapley Additive exPlanations), to provide insights into the reasoning behind the predictions made by the model. This can enhance user trust and facilitate effective human-in-the-loop decision-making.

b. **Visualizations and Interpretable Outputs:** Develop visualizations and interpretable outputs to assist users appreciate the model's predictions and identify the key factors contributing to the detected anomalies.

iv. **Real-time Implementation and Deployment:**

a. **Develop Real-time Systems:** Develop and evaluate real-time implementations of the proposed model to enable timely anomaly detection and response in operational environments. This will require optimizing the model for low latency and high throughput.

b. **Edge Computing:** Explore the deployment of the model on edge devices, such as edge servers and smart meters, to enable decentralized and real-time anomaly detection at the edge of the network.

v. **Advanced Ensemble Methods:**

a. **Investigate Novel Ensemble Techniques:** Explore the efficacy of more innovative ensemble methods, such as Bayesian model averaging, stacking, and boosting to increase the performance and robustness of anomaly detection models.

b. **Explore Heterogeneous Ensembles:** Investigate the probability of incorporating other machine learning models, such as isolation forests and one-class support vector machines, into the ensemble to further enhance diversity and improve overall performance.

vi. **Addressing Data Sparsity and Concept Drift:**

a. **Develop Techniques for Handling Data Sparsity:** Investigate techniques for handling data sparsity, such as domain adaptation, transfer learning, and data

augmentation, to improve the model's performance in scenarios with limited data availability.

b. **Address Concept Drift:** Explore techniques for addressing concept drift, that signifies temporal shifts in the data distribution. This is important for maintaining the model's effectiveness and accuracy in dynamic environments, particularly involving time-series data.

The identified areas for future research provide a structured roadmap for advancing the development of more effective anomaly detection systems in the context of energy consumption data. By systematically addressing these challenges and exploring the proposed research directions, future studies can contribute to the creation of more efficient, reliable, and sustainable energy systems through enhanced anomaly detection capabilities.

REFERENCES

[1] 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.

[2] T. Hu, Q. Guo, X. Shen, H. Sun, R. Wu, and H. Xi, 'Utilizing Unlabeled Data to Detect Electricity Fraud in AMI: A Semisupervised Deep Learning Approach', IEEE Trans. Neural Netw. Learning Syst., vol. 30, no. 11, pp. 3287–3299, Nov. 2019, doi: 10.1109/TNNLS.2018.2890663.

[3] R. Razavi-Far, M. Farajzadeh-Zanjani, B. Wang, M. Saif, and S. Chakrabarti, 'Imputation-based Ensemble Techniques for Class Imbalance Learning', IEEE Trans. Knowl. Data Eng., pp. 1–1, 2019, doi: 10.1109/TKDE.2019.2951556.

[4] X. Wang, H. Wang, B. Bhandari, and L. Cheng, 'AI-Empowered Methods for Smart Energy Consumption: A Review of Load Forecasting, Anomaly Detection and Demand Response', *Int. J. of Precis. Eng. and Manuf.-Green Tech.*, vol. 11, no. 3, pp. 963–993, May 2024, doi: 10.1007/s40684-023-00537-0.

[5] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, 'Imbalanced Data Problem in Machine Learning: A Review', *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[6] S. Maheshwari, R. C. Jain, and R. S. Jadon, 'A Review on Class Imbalance Problem: Analysis and Potential Solutions', *IJCSI*, vol. 14, no. 6, pp. 43–51, Nov. 2017, doi: 10.20943/01201706.4351.

[7] 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.

[8] S. K. Gunturi and D. Sarkar, 'Ensemble machine learning models for the detection of energy theft', *Electric Power Systems Research*, vol. 192, p. 106904, Mar. 2021, doi: 10.1016/j.epsr.2020.106904.

[9] J. Pereira and M. Silveira, 'Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention', in *2018*

17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL: IEEE, Dec. 2018, pp. 1275–1282. doi: 10.1109/ICMLA.2018.00207.

[10] S. Wang, G. Wang, Q. Fu, Y. Song, and J. Liu, 'IH-TCGAN: Time-Series Conditional Generative Adversarial Network with Improved Hausdorff Distance for Synthesizing Intention Recognition Data', *Entropy*, vol. 25, no. 5, p. 781, May 2023, doi: 10.3390/e25050781.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 'SMOTE: Synthetic Minority Over-sampling Technique', *jair*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[12] A. Telikani, A. H. Gandomi, K.-K. R. Choo, and J. Shen, 'A Cost-Sensitive Deep Learning-Based Approach for Network Traffic Classification', *IEEE Trans. Netw. Serv. Manage.*, vol. 19, no. 1, pp. 661–670, Mar. 2022, doi: 10.1109/TNSM.2021.3112283.

[13] H. Qin, H. Zhou, and J. Cao, 'Imbalanced learning algorithm based intelligent abnormal electricity consumption detection', *Neurocomputing*, vol. 402, pp. 112–123, Aug. 2020, doi: 10.1016/j.neucom.2020.03.085.

[14] M. Kardi, T. AlSkaif, B. Tekinerdogan, and J. P. S. Catalao, 'Anomaly Detection in Electricity Consumption Data using Deep Learning', in *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*, Bari, Italy: IEEE, Sep. 2021, pp. 1–6. doi: 10.1109/EEEIC/ICPSEurope51590.2021.9584650.

[15] S.-C. Yip, W.-N. Tan, C. Tan, M.-T. Gan, and K. Wong, 'An anomaly detection framework for identifying energy theft and defective meters in smart grids', *International Journal of Electrical Power & Energy Systems*, vol. 101, pp. 189–203, Oct. 2018, doi: 10.1016/j.ijepes.2018.03.025.

[16] L. Zhang, L. Wan, Y. Xiao, S. Li, and C. Zhu, 'Anomaly Detection method of Smart Meters data based on GMM-LDA clustering feature Learning and PSO Support Vector Machine', in *2019 IEEE Sustainable Power and Energy*

Conference (iSPEC), Beijing, China: IEEE, Nov. 2019, pp. 2407–2412. doi: 10.1109/iSPEC48194.2019.8974989.

[17] F. Harrou, A. Dairi, B. Taghezouit, and Y. Sun, 'An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine', *Solar Energy*, vol. 179, pp. 48–58, Feb. 2019, doi: 10.1016/j.solener.2018.12.045.

[18] A. Amara Korba and N. El Islem Karabadji, 'Smart Grid Energy Fraud Detection Using SVM', in *2019 International Conference on Networking and Advanced Systems (ICNAS)*, Annaba, Algeria: IEEE, Jun. 2019, pp. 1–6. doi: 10.1109/ICNAS.2019.8807832.

[19] S. Chen, X. Wu, and R. Zhang, 'A Novel Twin Support Vector Machine for Binary Classification Problems', *Neural Process Lett*, vol. 44, no. 3, pp. 795–811, Dec. 2016, doi: 10.1007/s11063-016-9495-0.

[20] J. Zhang, H. Zhang, S. Ding, and X. Zhang, 'Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means', *Front. Energy Res.*, vol. 9, p. 779587, Oct. 2021, doi: 10.3389/fenrg.2021.779587.

[21] M. Ahmed, R. Seraj, and S. M. S. Islam, 'The k-means Algorithm: A Comprehensive Survey and Performance Evaluation', *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

[22] W. Mao, X. Cao, Q. Zhou, T. Yan, and Y. Zhang, 'Anomaly Detection for Power Consumption Data based on Isolated Forest', in *2018 International Conference on Power System Technology (POWERCON)*, Guangzhou, China: IEEE, Nov. 2018, pp. 4169–4174. doi: 10.1109/POWERCON.2018.8602251.

[23] T. Mendes, P. J. S. Cardoso, J. Monteiro, and J. Raposo, 'Anomaly Detection of Consumption in Hotel Units: A Case Study Comparing Isolation Forest and Variational Autoencoder Algorithms', *Applied Sciences*, vol. 13, no. 1, p. 314, Dec. 2022, doi: 10.3390/app13010314.

[24] X. Wan, X. Zhang, and L. Liu, 'An Improved VGG19 Transfer Learning Strip Steel Surface Defect Recognition Deep Neural Network Based on Few Samples

and Imbalanced Datasets', *Applied Sciences*, vol. 11, no. 6, p. 2606, Mar. 2021, doi: 10.3390/app11062606.

[25] Y. Himeur, A. Alsalemi, F. Bensaali, and A. Amira, 'Detection of Appliance-Level Abnormal Energy Consumption in Buildings Using Autoencoders and Micro-moments', in *Proceedings of the 5th International Conference on Big Data and Internet of Things*, vol. 489, M. Lazaar, C. Duvallet, A. Touhafi, and M. Al Achhab, Eds., in Lecture Notes in Networks and Systems, vol. 489. , Cham: Springer International Publishing, 2022, pp. 179–193. doi: 10.1007/978-3-031-07969-6_14.

[26] J. P. Shaffer, 'Multiple Hypothesis Testing', *Annu. Rev. Psychol.*, vol. 46, no. 1, pp. 561–584, Jan. 1995, doi: 10.1146/annurev.ps.46.020195.003021.

[27] D. Elreedy, A. F. Atiya, and F. Kamalov, 'A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning', *Mach Learn*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[28] T. Pan, W. Pedrycz, J. Yang, and J. Wang, 'An improved generative adversarial network to oversample imbalanced datasets', *Engineering Applications of Artificial Intelligence*, vol. 132, p. 107934, Jun. 2024, doi: 10.1016/j.engappai.2024.107934.

[29] M. Emadaleslami and M. Haghifam, 'A Machine Learning Approach to Detect Energy Fraud in Smart Distribution Network', *IJSEE*, vol. 10, no. 02, Jun. 2021, doi: 10.30495/ijsee.2021.683248.

[30] Z. Wu, W. Lin, and Y. Ji, 'An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics', *IEEE Access*, vol. 6, pp. 8394–8402, 2018, doi: 10.1109/ACCESS.2018.2807121.

[31] R. Blagus and L. Lusa, 'SMOTE for high-dimensional class-imbalanced data', *BMC Bioinformatics*, vol. 14, no. 1, p. 106, Dec. 2013, doi: 10.1186/1471-2105-14-106.

[32] S. Kumari, C. Prabha, A. Karim, Md. M. Hassan, and S. Azam, 'A Comprehensive Investigation of Anomaly Detection Methods in Deep Learning

and Machine Learning: 2019–2023', *IET Information Security*, vol. 2024, no. 1, p. 8821891, Jan. 2024, doi: 10.1049/2024/8821891.

[33] C. Raghavendra and S. Chawla, 'Deep learning for anomaly detection: A survey', *arXiv preprint arXiv*, 2019.

[34] M. Ahmed, J. Chen, E. Akpaku, R. N. A. Sosu, and A. Latif, 'DELM: Deep Ensemble Learning Model for Anomaly Detection in Malicious Network Traffic-based Adaptive Feature Aggregation and Network Optimization', *ACM Trans. Priv. Secur.*, vol. 27, no. 4, pp. 1–36, Nov. 2024, doi: 10.1145/3690637.

[35] B. Azhari Shouni, S. Rahayu Selamat, Z. Zainal Abidin, and R. Wahyudi, 'Data Generative Model to Detect the Anomalies for IDS Imbalance CICIDS2017 Dataset', *TEM Journal*, vol. 12, no. 1, 2023.

[36] B. Xu and G. Yang, 'Interpretability research of deep learning: A literature survey', *Information Fusion*, vol. 115, p. 102721, Mar. 2025, doi: 10.1016/j.inffus.2024.102721.

[37] B. Gao *et al.*, 'Enhancing anomaly detection accuracy and interpretability in low-quality and class imbalanced data: A comprehensive approach', *Applied Energy*, vol. 353, p. 122157, Jan. 2024, doi: 10.1016/j.apenergy.2023.122157.

[38] 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.

[39] A. Ambat and J. Sahoo, 'Anomaly detection and prediction of energy consumption for smart homes using machine learning', *ETRI Journal*, p. etrij.2023-0155, Sep. 2024, doi: 10.4218/etrij.2023-0155.

[40] F. Fadli, Y. Himeur, M. Elnour, and A. Amira, 'Unveiling Hidden Energy Anomalies: Harnessing Deep Learning to Optimize Energy Management in Sports Facilities', 2024, *arXiv*. doi: 10.48550/ARXIV.2402.08742.

[41] R. L. Smith, 'EXTREME VALUE THEORY', in *in Handbook of applicable mathematics*, vol. 7, 1990, pp. 437–471.

[42] A. Ferreira and L. Haan, 'Extreme value theory', in *An introduction*, vol. 3, 2006.

[43] J. Galambos, ‘Extreme Value Theory for Applications’, in *Extreme Value Theory and Applications*, J. Galambos, J. Lechner, and E. Simiu, Eds., Boston, MA: Springer US, 1994, pp. 1–14. doi: 10.1007/978-1-4613-3638-9_1.

[44] J. Beirlant, Y. Goegebeur, and J. L. Teugels, *Statistics of extremes: theory and applications*. John Wiley & Sons., 2006.

[45] I. Markovsky and F. Dörfler, ‘Behavioral systems theory in data-driven analysis, signal processing, and control’, *Annual Reviews in Control*, vol. 52, pp. 42–64, 2021, doi: 10.1016/j.arcontrol.2021.09.005.

[46] F. Dörfler, Z. He, G. Belgioioso, S. Bolognani, J. Lygeros, and M. Muehlebach, ‘Toward a Systems Theory of Algorithms’, *IEEE Control Syst. Lett.*, vol. 8, pp. 1198–1210, 2024, doi: 10.1109/LCSYS.2024.3406943.

[47] F. Hasselman and A. M. T. Bosman, ‘Studying Complex Adaptive Systems With Internal States: A Recurrence Network Approach to the Analysis of Multivariate Time-Series Data Representing Self-Reports of Human Experience’, *Front. Appl. Math. Stat.*, vol. 6, p. 9, Apr. 2020, doi: 10.3389/fams.2020.00009.

[48] Å. Karl Johan and M. Richard, *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2021.

[49] A. Akundi and E. Smith, ‘Quantitative Characterization of Complex Systems—An Information Theoretic Approach’, *ASI*, vol. 4, no. 4, p. 99, Dec. 2021, doi: 10.3390/asi4040099.

[50] C. Gershenson, D. Polani, and G. Martius, ‘Editorial: Complexity and Self-Organization’, *Front. Robot. AI*, vol. 8, p. 668305, Mar. 2021, doi: 10.3389/frobt.2021.668305.

[51] M. Vidyasagar, ‘Learning Theory: the Probably Approximately Correct Framework’, in *Encyclopedia of Systems and Control*, J. Baillieul and T. Samad, Eds., Cham: Springer International Publishing, 2021, pp. 1103–1107. doi: 10.1007/978-3-030-44184-5_227.

[52] M. Merlin and B. Clarke, ‘Model selection via the VC dimension’, *Journal of Machine Learning Research*, vol. 20, no. 88, pp. 1–26, 2019.

[53] X. Guan and H. Burton, ‘Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications’, *Structures*, vol. 46, pp. 17–30, Dec. 2022, doi: 10.1016/j.istruc.2022.10.004.

[54] L. Zhang, G. Karakasidis, A. Odnoblyudova, L. Dogruel, Y. Tian, and A. Jung, ‘Explainable empirical risk minimization’, *Neural Comput & Applic*, vol. 36, no. 8, pp. 3983–3996, Mar. 2024, doi: 10.1007/s00521-023-09269-3.

[55] R. Ballester, X. A. Clemente, C. Casacuberta, M. Madadi, C. A. Corneanu, and S. Escalera, ‘Predicting the generalization gap in neural networks using topological data analysis’, *Neurocomputing*, vol. 596, p. 127787, Sep. 2024, doi: 10.1016/j.neucom.2024.127787.

[56] T. Ciprian-Octavian and C. Adrian Leordeanu, ‘Classification of an imbalanced data set using decision tree algorithms’, *Univ. Politech. Bucharest Sci. Bull. Ser. C Electr. Eng. Comput. Sci*, vol. 79, pp. 69–84, 2017.

[57] A. S. More and D. P. Rana, ‘Review of random forest classification techniques to resolve data imbalance’, in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Aurangabad: IEEE, Oct. 2017, pp. 72–78. doi: 10.1109/ICISIM.2017.8122151.

[58] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, ‘An Empirical Study of Learning from Imbalanced Data Using Random Forest’, in *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, Patras, Greece: IEEE, Oct. 2007, pp. 310–317. doi: 10.1109/ICTAI.2007.46.

[59] S. B. S. Lai, N. H. N. Binti Md Shahri, M. B. Mohamad, H. A. B. A. Rahman, and A. B. Rambli, ‘Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data’, *ms*, vol. 9, no. 3, pp. 379–385, May 2021, doi: 10.13189/ms.2021.090320.

[60] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, ‘An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset’, *Computer Networks*, vol. 177, p. 107315, Aug. 2020, doi: 10.1016/j.comnet.2020.107315.

[61] 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 Trans. Ind. Inf.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018, doi: 10.1109/TII.2017.2785963.

[62] A. Radford, L. Metz, and S. Chintala, ‘Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks’, Jan. 07, 2016, *arXiv*: arXiv:1511.06434. doi: 10.48550/arXiv.1511.06434.

[63] C. Xu and H. Chen, ‘A hybrid data mining approach for anomaly detection and evaluation in residential buildings energy data’, *Energy and Buildings*, vol. 215, p. 109864, May 2020, doi: 10.1016/j.enbuild.2020.109864.

[64] W. Dai, X. Liu, A. Heller, and P. S. Nielsen, ‘Smart Meter Data Anomaly Detection Using Variational Recurrent Autoencoders with Attention’, in *Intelligent Technologies and Applications*, vol. 1616, F. Sanfilippo, O.-C. Granmo, S. Y. Yayilgan, and I. S. Bajwa, Eds., in Communications in Computer and Information Science, vol. 1616. , Cham: Springer International Publishing, 2022, pp. 311–324. doi: 10.1007/978-3-031-10525-8_25.

[65] Y. Zhao, Z. Nasrullah, and Z. Li, ‘PyOD: A Python Toolbox for Scalable Outlier Detection’, Jun. 10, 2019, *arXiv*: arXiv:1901.01588. doi: 10.48550/arXiv.1901.01588.

[66] 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.

[67] L. Yulong and J. Lu, ‘A universal approximation theorem of deep neural networks for expressing probability distributions’, *Advances in neural information processing systems*, vol. 33, pp. 3094–3105, 2020.

[68] H. Boris and D. Rolnick, ‘Complexity of linear regions in deep networks’, *International Conference on Machine Learning. PMLR*, 2019.

[69] C. Setareh, K. N. H, D. Rolnick, and van de P. M., ‘Understanding the evolution of linear regions in deep reinforcement learning’, *Advances in Neural Information Processing Systems*, vol. 35, pp. 10891–10903, 2022.

[70] L. Alzubaidi *et al.*, ‘Review of deep learning: concepts, CNN architectures, challenges, applications, future directions’, *J Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s4053`7-021-00444-8.

[71] K. A. Sankararaman, S. De, Z. Xu, W. R. Huang, and T. Goldstein, ‘The Impact of Neural Network Overparameterization on Gradient Confusion and Stochastic Gradient Descent’, *in Proceedings of the 37th International Conference on Machine Learning, PMLR 119, Online*, 2020.

[72] J. E. Dobson, ‘On reading and interpreting black box deep neural networks’, *Int J Digit Humanities*, vol. 5, no. 2–3, pp. 431–449, Nov. 2023, doi: 10.1007/s42803-023-00075-w.

[73] M. Dialameh, A. Hamzeh, H. Rahmani, S. Dialameh, and H. J. Kwon, ‘DL-Reg: A deep learning regularization technique using linear regression’, *Expert Systems with Applications*, vol. 247, p. 123182, Aug. 2024, doi: 10.1016/j.eswa.2024.123182.

[74] F. Févotte, A. Rappaport, and M. Vohralík, ‘Adaptive regularization, discretization, and linearization for nonsmooth problems based on primal–dual gap estimators’, *Computer Methods in Applied Mechanics and Engineering*, vol. 418, p. 116558, Jan. 2024, doi: 10.1016/j.cma.2023.116558.

[75] I. Salehin and D.-K. Kang, ‘A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain’, *Electronics*, vol. 12, no. 14, p. 3106, Jul. 2023, doi: 10.3390/electronics12143106.

[76] P. Soltanzadeh and M. Hashemzadeh, ‘RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem’, *Information Sciences*, vol. 542, pp. 92–111, Jan. 2021, doi: 10.1016/j.ins.2020.07.014.

[77] D. S. Sisodia and U. Verma, ‘Distinct Multiple Learner-Based Ensemble SMOTEBagging (ML-ESB) Method for Classification of Binary Class

Imbalance Problems', *IJTech*, vol. 10, no. 4, p. 721, Jul. 2019, doi: 10.14716/ijtech.v10i4.1743.

[78] T. Hasanin and T. Khoshgoftaar, 'The Effects of Random Undersampling with Simulated Class Imbalance for Big Data', in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, Salt Lake City, UT: IEEE, Jul. 2018, pp. 70–79. doi: 10.1109/IRI.2018.00018.

[79] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, 'A survey on addressing high-class imbalance in big data', *J Big Data*, vol. 5, no. 1, p. 42, Dec. 2018, doi: 10.1186/s40537-018-0151-6.

[80] J. M. Johnson and T. M. Khoshgoftaar, 'Survey on deep learning with class imbalance', *J Big Data*, vol. 6, no. 1, p. 27, Dec. 2019, doi: 10.1186/s40537-019-0192-5.

[81] M. Buda, A. Maki, and M. A. Mazurowski, 'A systematic study of the class imbalance problem in convolutional neural networks', *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[82] A. K. Menon and R. C. Williamson, 'The institutions of imbalance: a study of class imbalance learning', *ICML*, 2018.

[83] L. Liu, X. Wu, S. Li, Y. Li, S. Tan, and Y. Bai, 'Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection', *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 82, Dec. 2022, doi: 10.1186/s12911-022-01821-w.

[84] M. Temraz and M. T. Keane, 'Solving the class imbalance problem using a counterfactual method for data augmentation', *Machine Learning with Applications*, vol. 9, p. 100375, Sep. 2022, doi: 10.1016/j.mlwa.2022.100375.

[85] Z. Wu, H. Wang, J. Guo, Q. Yang, and J. Shao, 'Learning evolving prototypes for imbalanced data stream classification with limited labels', *Information Sciences*, vol. 679, p. 120979, Sep. 2024, doi: 10.1016/j.ins.2024.120979.

[86] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, 'The receiver operating characteristic curve accurately assesses imbalanced datasets', *Patterns*, vol. 5, no. 6, 2024.

[87] A. N. Alkuwari, S. Al-Kuwari, and M. Qaraqe, ‘Anomaly Detection in Smart Grids: A Survey From Cybersecurity Perspective’, in *2022 3rd International Conference on Smart Grid and Renewable Energy (SGRE)*, Doha, Qatar: IEEE, Mar. 2022, pp. 1–7. doi: 10.1109/SGRE53517.2022.9774221.

[88] J. E. Zhang, D. Wu, and B. Boulet, ‘Time Series Anomaly Detection for Smart Grids: A Survey’, in *2021 IEEE Electrical Power and Energy Conference (EPEC)*, Toronto, ON, Canada: IEEE, Oct. 2021, pp. 125–130. doi: 10.1109/EPEC52095.2021.9621752.

[89] W. Ju *et al.*, ‘A Comprehensive Survey on Deep Graph Representation Learning’, *Neural Networks*, vol. 173, p. 106207, May 2024, doi: 10.1016/j.neunet.2024.106207.

[90] K. T.N, *Deep learning with graph-structured representations*. Amsterdam, The Netherlands: Informatics Institute, 2020.

[91] J. Liu, K. Song, M. Feng, Y. Yan, Z. Tu, and L. Zhu, ‘Semi-supervised anomaly detection with dual prototypes autoencoder for industrial surface inspection’, *Optics and Lasers in Engineering*, vol. 136, p. 106324, Jan. 2021, doi: 10.1016/j.optlaseng.2020.106324.

[92] R. ELhadad, Y.-F. Tan, and W.-N. Tan, ‘Anomaly Prediction in Electricity Consumption Using a Combination of Machine Learning Techniques’, *IJTech*, vol. 13, no. 6, p. 1317, Nov. 2022, doi: 10.14716/ijtech.v13i6.5931.

[93] S. Sørbø and M. Ruocco, ‘Navigating the metric maze: a taxonomy of evaluation metrics for anomaly detection in time series’, *Data Min Knowl Disc*, vol. 38, no. 3, pp. 1027–1068, May 2024, doi: 10.1007/s10618-023-00988-8.

[94] Ö. Sen, S. Glomb, M. Henze, and A. Ulbig, ‘Benchmark Evaluation of Anomaly-Based Intrusion Detection Systems in the Context of Smart Grids’, in *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, Grenoble, France: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ISGTEUROPE56780.2023.10407262.

[95] G. Sri Vidhya and R. Nagarajan, ‘A novel bidirectional LSTM model for network intrusion detection in SDN-IoT network’, *Computing*, vol. 106, no. 8, pp. 2613–2642, Aug. 2024, doi: 10.1007/s00607-024-01295-w.

[96] P. Wang, K. Wang, Y. Song, and X. Wang, ‘AutoLDT: a lightweight spatio-temporal decoupling transformer framework with AutoML method for time series classification’, *Sci Rep*, vol. 14, no. 1, p. 29801, Nov. 2024, doi: 10.1038/s41598-024-81000-1.

[97] S. Tuli, G. Casale, and N. R. Jennings, ‘TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data’, 2022, *arXiv*. doi: 10.48550/ARXIV.2201.07284.

[98] R. Zaccagnino, A. Cirillo, A. Guarino, N. Lettieri, D. Malandrino, and G. Zaccagnino, ‘Towards a Geometric Deep Learning-Based Cyber Security: Network System Intrusion Detection Using Graph Neural Networks’:, in *Proceedings of the 20th International Conference on Security and Cryptography*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2023, pp. 394–401. doi: 10.5220/0012085700003555.

[99] C. Landin, J. Liu, K. Katsarou, and S. Tahvili, ‘Time Series Anomaly Detection using Convolutional Neural Networks in the Manufacturing Process of RAN’, in *2023 IEEE International Conference On Artificial Intelligence Testing (AITest)*, Athens, Greece: IEEE, Jul. 2023, pp. 90–98. doi: 10.1109/AITest58265.2023.00023.

[100] F. Laghrissi, S. Douzi, K. Douzi, and B. Hssina, ‘Intrusion detection systems using long short-term memory (LSTM)’, *J Big Data*, vol. 8, no. 1, p. 65, Dec. 2021, doi: 10.1186/s40537-021-00448-4.

[101] Y. Imrana, Y. Xiang, L. Ali, and Z. Abdul-Rauf, ‘A bidirectional LSTM deep learning approach for intrusion detection’, *Expert Systems with Applications*, vol. 185, p. 115524, Dec. 2021, doi: 10.1016/j.eswa.2021.115524.

[102] Y. Yang, S. Tu, R. Hashim Ali, H. Alasmary, M. Waqas, and M. Nouman Amjad, ‘Intrusion Detection Based on Bidirectional Long Short-Term Memory with Attention Mechanism’, *Computers, Materials & Continua*, vol. 74, no. 1, pp. 801–815, 2023, doi: 10.32604/cmc.2023.031907.

[103] A. Siva Kumar, S. Raja, N. Pritha, H. Raviraj, R. Babitha Lincy, and J. Jency Rubia, ‘An adaptive transformer model for anomaly detection in wireless sensor networks in real-time’, *Measurement: Sensors*, vol. 25, p. 100625, Feb. 2023, doi: 10.1016/j.measen.2022.100625.

[104] S. Fu *et al.*, ‘A time series anomaly detection method based on series-parallel transformers with spatial and temporal association discrepancies’, *Information Sciences*, vol. 657, p. 119978, Feb. 2024, doi: 10.1016/j.ins.2023.119978.

[105] L. Su, X. Zuo, R. Li, X. Wang, H. Zhao, and B. Huang, ‘A systematic review for transformer-based long-term series forecasting’, *Artif Intell Rev*, vol. 58, no. 3, p. 80, Jan. 2025, doi: 10.1007/s10462-024-11044-2.

[106] W. Lim, K. S. C. Yong, B. T. Lau, and C. C. L. Tan, ‘Future of generative adversarial networks (GAN) for anomaly detection in network security: A review’, *Computers & Security*, vol. 139, p. 103733, Apr. 2024, doi: 10.1016/j.cose.2024.103733.

[107] M. T. Ribeiro, S. Singh, and C. Guestrin, ‘“Why Should I Trust You?”: Explaining the Predictions of Any Classifier’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[108] A. M. Salih *et al.*, ‘A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME’, *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, Jan. 2025, doi: 10.1002/aisy.202400304.

[109] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, ‘Explanation of machine learning models using shapley additive explanation and application for real data in hospital’, *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106584, Feb. 2022, doi: 10.1016/j.cmpb.2021.106584.

[110] A. Houdou *et al.*, ‘Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review’, *Aerosol Air Qual. Res.*, vol. 24, no. 1, p. 230151, 2024, doi: 10.4209/aaqr.230151.

[111] L. A. Cox, ‘What is an exposure-response curve?’, *Global Epidemiology*, vol. 6, p. 100114, Dec. 2023, doi: 10.1016/j.gloepi.2023.100114.

[112] A. Li, F. Xiao, C. Zhang, and C. Fan, ‘Attention-based interpretable neural network for building cooling load prediction’, *Applied Energy*, vol. 299, p. 117238, Oct. 2021, doi: 10.1016/j.apenergy.2021.117238.

[113] B. Kotipalli, ‘The Role of Attention Mechanisms in Enhancing Transparency and Interpretability of Neural Network Models in Explainable AI’, *Digital Commons at Harrisburg University*, 2024.

[114] T. Danesh, R. Ouaret, P. Floquet, and S. Negny, ‘Hybridization of model-specific and model-agnostic methods for interpretability of Neural network predictions: Application to a power plant’, *Computers & Chemical Engineering*, vol. 176, p. 108306, Aug. 2023, doi: 10.1016/j.compchemeng.2023.108306.

[115] C. Molnar *et al.*, ‘General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models’, in *xxAI - Beyond Explainable AI*, vol. 13200, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., in Lecture Notes in Computer Science, vol. 13200. , Cham: Springer International Publishing, 2022, pp. 39–68. doi: 10.1007/978-3-031-04083-2_4.

[116] S. Touzani, J. Granderson, and S. Fernandes, ‘Gradient boosting machine for modeling the energy consumption of commercial buildings’, *Energy and Buildings*, vol. 158, pp. 1533–1543, Jan. 2018, doi: 10.1016/j.enbuild.2017.11.039.

[117] Z. Qu, H. Li, Y. Wang, J. Zhang, A. Abu-Siada, and Y. Yao, ‘Detection of Electricity Theft Behavior Based on Improved Synthetic Minority Oversampling Technique and Random Forest Classifier’, *Energies*, vol. 13, no. 8, p. 2039, Apr. 2020, doi: 10.3390/en13082039.

[118] A. Mohammed and R. Kora, ‘A comprehensive review on ensemble deep learning: Opportunities and challenges’, *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.

[119] L. Alzubaidi *et al.*, 'MEFF – A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging', *Intelligent Systems with Applications*, vol. 22, p. 200355, Jun. 2024, doi: 10.1016/j.iswa.2024.200355.

[120] Y. Kwon and J. Zou, 'Data-OOB: Out-of-bag Estimate as a Simple and Efficient Data Value', *in Proceedings of the 40th International Conference on Machine Learning (PMLR 202),Honolulu, Hawaii, USA*, 2023.

[121] J.-Y. Kim and S.-B. Cho, 'Evolutionary Optimization of Hyperparameters in Deep Learning Models', in *2019 IEEE Congress on Evolutionary Computation (CEC)*, Wellington, New Zealand: IEEE, Jun. 2019, pp. 831–837. doi: 10.1109/CEC.2019.8790354.

[122] M. Wojciuk, Z. Swiderska-Chadaj, K. Siwek, and A. Gertych, 'Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization', *Heliyon*, vol. 10, no. 5, p. e26586, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26586.

[123] M. A. K. Raiaan *et al.*, 'A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks', *Decision Analytics Journal*, vol. 11, p. 100470, Jun. 2024, doi: 10.1016/j.dajour.2024.100470.

[124] L. Krstic, M. Ivanovic, V. Simic, and B. Stojanovic, 'Evolutionary approach for composing a thoroughly optimized ensemble of regression neural networks', *Egyptian Informatics Journal*, vol. 28, p. 100581, Dec. 2024, doi: 10.1016/j.eij.2024.100581.

[125] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, 'Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data', *J Big Data*, vol. 11, no. 1, p. 7, Jan. 2024, doi: 10.1186/s40537-023-00857-7.

[126] I. D. Mienye and Y. Sun, 'Performance analysis of cost-sensitive learning methods with application to imbalanced medical data', *Informatics in Medicine Unlocked*, vol. 25, p. 100690, 2021, doi: 10.1016/j.imu.2021.100690.

[127] Z. Ren *et al.*, 'Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data', *Knowledge-Based Systems*, vol. 241, p. 108296, Apr. 2022, doi: 10.1016/j.knosys.2022.108296.

[128] L. Zhang, G. Sheng, H. Hou, N. Zhou, and X. Jiang, 'An adaptive fault diagnosis method of power transformers based on combining oversampling and cost-sensitive learning', *IET Smart Grid*, vol. 4, no. 6, pp. 623–635, Dec. 2021, doi: 10.1049/stg2.12044.

[129] Z. Chen, J. Duan, L. Kang, and G. Qiu, 'A hybrid data-level ensemble to enable learning from highly imbalanced dataset', *Information Sciences*, vol. 554, pp. 157–176, Apr. 2021, doi: 10.1016/j.ins.2020.12.023.

[130] R. A. Sowah *et al.*, 'HCBST: An Efficient Hybrid Sampling Technique for Class Imbalance Problems', *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 3, pp. 1–37, Jun. 2022, doi: 10.1145/3488280.

[131] T. Hugo, F. Popineau, A. Rimmel, B.-L. Doan, and F. Daniel, 'Comparative evaluation of anomaly detection methods for fraud detection in online credit card payments', *Springer Nature Singapore*, pp. 37–50, 2024.

[132] J. Kong, W. Kowalczyk, S. Menzel, and T. Bäck, 'Improving Imbalanced Classification by Anomaly Detection', in *Parallel Problem Solving from Nature – PPSN XVI*, vol. 12269, T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, and H. Trautmann, Eds., in Lecture Notes in Computer Science, vol. 12269. , Cham: Springer International Publishing, 2020, pp. 512–523. doi: 10.1007/978-3-030-58112-1_35.

[133] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza, 'General Performance Score for classification problems', *Appl Intell*, vol. 52, no. 10, pp. 12049–12063, Aug. 2022, doi: 10.1007/s10489-021-030417.

[134] Z. Qu, H. Liu, Z. Wang, J. Xu, P. Zhang, and H. Zeng, 'A combined genetic optimization with AdaBoost ensemble model for anomaly detection in buildings electricity consumption', *Energy and Buildings*, vol. 248, p. 111193, Oct. 2021, doi: 10.1016/j.enbuild.2021.111193.

[135] N. Javaid, N. Jan, and M. U. Javed, ‘An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids’, *Journal of Parallel and Distributed Computing*, vol. 153, pp. 44–52, Jul. 2021, doi: 10.1016/j.jpdc.2021.03.002.

[136] R. A. Gomez and F. A. Llaugel, ‘Electrical Network Fraud Detection Using a CNN+LSTM Model’, *IJAEMR*, vol. 08, no. 05, pp. 54–64, 2023, doi: 10.51505/ijaemr.2023.8505.

[137] Z. A. Khan, M. Adil, N. Javaid, M. N. Saqib, M. Shafiq, and J.-G. Choi, ‘Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data’, *Sustainability*, vol. 12, no. 19, p. 8023, Sep. 2020, doi: 10.3390/su12198023.

[138] L. J. Lepolesa, S. Achari, and L. Cheng, ‘Electricity Theft Detection in Smart Grids Based on Deep Neural Network’, *IEEE Access*, vol. 10, pp. 39638–39655, 2022, doi: 10.1109/ACCESS.2022.3166146.

[139] J. Lee, Y. C. Lee, and J. T. Kim, ‘Fault detection based on one-class deep learning for manufacturing applications limited to an imbalanced database’, *Journal of Manufacturing Systems*, vol. 57, pp. 357–366, Oct. 2020, doi: 10.1016/j.jmsy.2020.10.013.

[140] L. Zhang, X. Chen, C. Zhang, L. Zhang, and B. Zou, ‘Intelligent Anomaly Detection Method of Gateway Electrical Energy Metering Devices using Deep Learning’, *IJACSA*, vol. 14, no. 7, 2023, doi: 10.14569/IJACSA.2023.0140793.

[141] A. Castellani, S. Schmitt, and S. Squartini, ‘Real-World Anomaly Detection by Using Digital Twin Systems and Weakly Supervised Learning’, *IEEE Trans. Ind. Inf.*, vol. 17, no. 7, pp. 4733–4742, Jul. 2021, doi: 10.1109/TII.2020.3019788.

[142] D. Cao, D. Liu, X. Ren, and N. Ma, ‘Self-Adaption AAE-GAN for Aluminum Electrolytic Cell Anomaly Detection’, *IEEE Access*, vol. 9, pp. 100991–101002, 2021, doi: 10.1109/ACCESS.2021.3097116.

[143] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, ‘Hybrid Deep Neural Networks for Detection of Non-Technical Losses in

Electricity Smart Meters', *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1254–1263, Mar. 2020, doi: 10.1109/TPWRS.2019.2943115.

[144] J. Davis and M. Goadrich, 'The relationship between Precision-Recall and ROC curves', in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[145] D. Chicco and G. Jurman, 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[146] W. Khreich, E. Granger, A. Miri, and R. Sabourin, 'Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs', *Pattern Recognition*, vol. 43, no. 8, pp. 2732–2752, Aug. 2010, doi: 10.1016/j.patcog.2010.03.006.

[147] I. Petrlik, P. Lezama, C. Rodriguez, R. Inquilla, J. E. Reyna-González, and R. Esparza, 'Electricity Theft Detection using Machine Learning', *IJACSA*, vol. 13, no. 12, 2022, doi: 10.14569/IJACSA.2022.0131251.

[148] J. Yao and M. Shepperd, 'Assessing software defection prediction performance: why using the Matthews correlation coefficient matters', in *Proceedings of the Evaluation and Assessment in Software Engineering*, Trondheim Norway: ACM, Apr. 2020, pp. 120–129. doi: 10.1145/3383219.3383232.

[149] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, 'Classification with class imbalance problem', *Int. J. Advance Soft Compu. Appl*, vol. 5, no. 3, Nov. 2013.

[150] J. Kua, M. B. Hossain, I. Natgunanathan, and Y. Xiang, 'Privacy Preservation in Smart Meters: Current Status, Challenges and Future Directions', *Sensors*, vol. 23, no. 7, p. 3697, Apr. 2023, doi: 10.3390/s23073697.

[151] M. de Medeiros *et al.*, 'Verified Foundations for Differential Privacy', 2024, *arXiv*. doi: 10.48550/ARXIV.2412.01671.

[152] P. Voigt and A. Von Dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Cham: Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-62328-8.

[153] Dr. V. P. S. , 'How can we manage biases in artificial intelligence systems – A systematic literature review', *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100165, Apr. 2023, doi: 10.1016/j.jjimei.2023.100165.

[154] C. C. PEREZ, *Invisible women: Data bias in a world designed for men.* Penguin Random House UK, 2019.

[155] E. Strubell, A. Ganesh, and A. McCallum, 'Energy and Policy Considerations for Modern Deep Learning Research', *AAAI*, vol. 34, no. 09, pp. 13693–13696, Apr. 2020, doi: 10.1609/aaai.v34i09.7123.

[156] N. A. Smuha, 'The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence', *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, Aug. 2019, doi: 10.9785/cri-2019-200402.

[157] S. P. Kaur, 'Variables in research', *Indian J. Res. Rep. Med. Sci*, vol. 3, no. 4, pp. 36–38, 2013.

[158] W. Zhu, N. Zeng, and N. Wang, 'Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations', *Nesug Proc Health Care Life Sci Balt Md*, vol. 19, no. 67, 2010.

[159] M. Ryo and M. C. Rillig, 'Statistically reinforced machine learning for nonlinear patterns and variable interactions', *Ecosphere*, vol. 8, no. 11, p. e01976, Nov. 2017, doi: 10.1002/ecs2.1976.

[160] J. W. Creswell and N. P. Cheryl, *Qualitative inquiry and research design: Choosing among five approaches.* Sage publications, 2016.

[161] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, 'Five Steps to Conducting a Systematic Review', *J R Soc Med*, vol. 96, no. 3, pp. 118–121, Mar. 2003, doi: 10.1177/014107680309600304.

[162] P. Järvinen, 'Mapping Research Questions to Research Methods', in *Advances in Information Systems Research, Education and Practice*, vol. 274, D. Avison,

G. M. Kasper, B. Pernici, I. Ramos, and D. Roode, Eds., in IFIP – The International Federation for Information Processing, vol. 274. , Boston, MA: Springer US, 2008, pp. 29–41. doi: 10.1007/978-0-387-09682-7-9_3.

[163] A. Soin, M. Hirschbeck, M. Verdon, and L. Manchikanti, 'A Pilot Study Implementing a Machine Learning Algorithm to Use Artificial Intelligence to Diagnose Spinal Conditions', *Pain Physician*, vol. 23, pp. 171–178, 2022.

[164] R. Buresh, E. Austin, and E. Craswell, 'Analytical methods in15N research', *Fertilizer Research*, vol. 3, no. 1, pp. 37–62, 1982, doi: 10.1007/BF01063408.

[165] M. M. Mohamad, N. L. Sulaiman, L. C. Sern, and K. M. Salleh, 'Measuring the Validity and Reliability of Research Instruments', *Procedia - Social and Behavioral Sciences*, vol. 204, pp. 164–171, Aug. 2015, doi: 10.1016/j.sbspro.2015.08.129.

[166] I. Etikan, 'Comparison of Convenience Sampling and Purposive Sampling', *AJTAS*, vol. 5, no. 1, p. 1, 2016, doi: 10.11648/j.ajtas.20160501.11.

[167] M. Valenti, R. Bonfigli, E. Principi, and A. S. Squartini, 'Exploiting the Reactive Power in Deep Neural Models for Non-Intrusive Load Monitoring', in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro: IEEE, Jul. 2018, pp. 1–8. doi: 10.1109/IJCNN.2018.8489271.

[168] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[169] M. S. Ebrahimi Shahabadi, H. Tabrizchi, M. Kuchaki Rafsanjani, B. B. Gupta, and F. Palmieri, 'A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems', *Technological Forecasting and Social Change*, vol. 169, p. 120796, Aug. 2021, doi: 10.1016/j.techfore.2021.120796.

[170] L. Minvielle, M. Atiq, S. Peignier, and M. Mougeot, 'Transfer Learning on Decision Tree with Class Imbalance', in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA: IEEE, Nov. 2019, pp. 1003–1010. doi: 10.1109/ICTAI.2019.00141.

[171] M. Salem, S. Taheri, and J. S. Yuan, ‘Anomaly Generation Using Generative Adversarial Networks in Host-Based Intrusion Detection’, in *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York City, NY, USA: IEEE, Nov. 2018, pp. 683–687. doi: 10.1109/UEMCON.2018.8796769.

[172] M. Tyagi, S. Roy, and V. Bansal, ‘Custom Weighted Balanced Loss function for Covid 19 Detection from an Imbalanced CXR Dataset’, in *2022 26th International Conference on Pattern Recognition (ICPR)*, Montreal, QC, Canada: IEEE, Aug. 2022, pp. 2707–2713. doi: 10.1109/ICPR56361.2022.9956580.

[173] W. Shi, R. Zhu, and S. Li, ‘Pairwise Adversarial Training for Unsupervised Class-imbalanced Domain Adaptation’, in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC USA: ACM, Aug. 2022, pp. 1598–1606. doi: 10.1145/3534678.3539243.

[174] S. Shukla and B. S. Raghuwanshi, ‘Online sequential class-specific extreme learning machine for binary imbalanced learning’, *Neural Networks*, vol. 119, pp. 235–248, Nov. 2019, doi: 10.1016/j.neunet.2019.08.018.

[175] L. I. S. Cln, ‘Data collection techniques a guide for researchers in humanities and education’, *International Research Journal of Computer Science and Information Systems (IRJCSIS)International Research Journal of Computer Science and Information Systems (IRJCSIS)*, vol. 2, no. 3, pp. 40–44, 2013.

[176] J. DeCoster, M. J. Banner, E. R. Smith, and G. R. Semin, ‘On the Inexplicability of the Implicit: Differences in the Information Provided by Implicit and Explicit Tests’, *Social Cognition*, vol. 24, no. 1, pp. 5–21, Feb. 2006, doi: 10.1521/soco.2006.24.1.5.

[177] J. F. Nunamaker, M. Chen, and T. D. M. Purdin, ‘Systems Development in Information Systems Research’, *Journal of Management Information Systems*, vol. 7, no. 3, pp. 89–106, Dec. 1990, doi: 10.1080/07421222.1990.11517898.

[178] S. M. Tabatabaie Nezhad, M. Nazari, and E. A. Gharavol, ‘A Novel DoS and DDoS Attacks Detection Algorithm Using ARIMA Time Series Model and

Chaotic System in Computer Networks', *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 700–703, Apr. 2016, doi: 10.1109/LCOMM.2016.2517622.

[179] J. Gu and S. Lu, 'An effective intrusion detection approach using SVM with naïve Bayes feature embedding', *Computers & Security*, vol. 103, p. 102158, Apr. 2021, doi: 10.1016/j.cose.2020.102158.

[180] M. Ahsan and K. E. Nygard, 'Convolutional Neural Networks with LSTM for Intrusion Detection', in *Proceedings of 35th International Conference on Computers and Their Applications*, 2020, pp. 69–79.

[181] 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, doi: 10.1016/j.prime.2022.100030.

[182] Y. Himeur, A. Alsalemi, F. Bensaali, and A. Amira, 'A Novel Approach for Detecting Anomalous Energy Consumption Based on Micro-Moments and Deep Neural Networks', *Cogn Comput*, vol. 12, no. 6, pp. 1381–1401, Nov. 2020, doi: 10.1007/s12559-020-09764-y.

[183] G. Huang, F. Wu, and C. Guo, 'Smart grid dispatch powered by deep learning: a survey', *Front Inform Technol Electron Eng*, vol. 23, no. 5, pp. 763–776, May 2022, doi: 10.1631/FITEE.2000719.

[184] H.-X. Gao, S. Kuenzel, and X.-Y. Zhang, 'A Hybrid ConvLSTM-Based Anomaly Detection Approach for Combating Energy Theft', *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022, doi: 10.1109/TIM.2022.3201569.

[185] Q. Huang *et al.*, 'A Novel Electricity Theft Detection Strategy Based on Dual-Time Feature Fusion and Deep Learning Methods', *Energies*, vol. 17, no. 2, p. 275, Jan. 2024, doi: 10.3390/en17020275.

[186] 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: IEEE, Jun. 2023, pp. 1–6. doi: 10.1109/VTC2023-Spring57618.2023.10200352.

[187] 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 Trans & Dist*, vol. 18, no. 3, pp. 413–445, Feb. 2024, doi: 10.1049/gtd2.13056.

[188] P. Jafarzadeh, J. Paalassalo, and o Eerola, ‘IOT-based household energy consumption prediction using machine learning’, in *Advances in Industrial Internet of Things, Engineering and Management*, Cham: Springer International Publishing., 2021, pp. 137-152.

[189] S. U. Khan, N. Khan, F. U. M. Ullah, M. J. Kim, M. Y. Lee, and S. W. Baik, ‘Towards intelligent building energy management: AI-based framework for power consumption and generation forecasting’, *Energy and Buildings*, vol. 279, p. 112705, Jan. 2023, doi: 10.1016/j.enbuild.2022.112705.

[190] M. El-Maraghy *et al.*, ‘Predicting energy consumption of mosque buildings during the operation stage using deep learning approach’, *Energy and Buildings*, vol. 303, p. 113829, Jan. 2024, doi: 10.1016/j.enbuild.2023.113829.

[191] 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.

[192] 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.

[193] 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: IEEE, Nov. 2022, pp. 1–5. doi: 10.1109/ASSIC55218.2022.10088334.

[194] J. Lu, Q. Zhang, Z. Yang, and M. Tu, ‘A hybrid model based on convolutional neural network and long short-term memory for short-term load forecasting’, in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, Atlanta, GA, USA: IEEE, Aug. 2019, pp. 1–5. doi: 10.1109/PESGM40551.2019.8973549.

[195] 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: IEEE, Sep. 2017, pp. 79–85. doi: 10.1109/ICACCI.2017.8125820.

[196] 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’, *Multimed Tools Appl*, vol. 80, no. 11, pp. 16979–16995, May 2021, doi: 10.1007/s11042-020-09406-3.

[197] 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, Apr. 2021, doi: 10.1016/j.ymssp.2020.107398.

[198] 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, no. 1, p. e0292140, Jan. 2024, doi: 10.1371/journal.pone.0292140.

[199] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, ‘Moving to a World Beyond “ $p < 0.05$”’, *The American Statistician*, vol. 73, no. sup1, pp. 1–19, Mar. 2019, doi: 10.1080/00031305.2019.1583913.

[200] R. Xia, Y. Gao, Y. Zhu, D. Gu, and J. Wang, ‘An Efficient Method Combined Data-Driven for Detecting Electricity Theft with Stacking Structure Based on Grey Relation Analysis’, *Energies*, vol. 15, no. 19, p. 7423, Oct. 2022, doi: 10.3390/en15197423.

[201] S. Aslam, H. Herodotou, S. M. Mohsin, N. Javaid, N. Ashraf, and S. Aslam, ‘A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids’, *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110992, Jul. 2021, doi: 10.1016/j.rser.2021.110992.

[202] F. Zhou *et al.*, ‘A Comprehensive Survey for Deep-Learning-Based Abnormality Detection in Smart Grids with Multimodal Image Data’, *Applied Sciences*, vol. 12, no. 11, p. 5336, May 2022, doi: 10.3390/app12115336.

APPENDICES

Appendix 1: Experimental Checklist

Item	Completed (Yes/No)	Comments
Data Preparation		Data Collection Data Cleaning Data preprocessing Data Splitting
Feature engineering		Feature selection Feature extraction Feature dimensionality reduction
Model Selection and Training		Model Selection Hyperparameter Tuning Cross-Validation Model Training process
Model Evaluation		Evaluation Metrics (e.g., accuracy, precision, recall, F1-score, AUC-ROC).

Appendix 2: Observation Checklist

Item	Completed (Yes/No)	Comments
Models selected		List of models selected
Training/validation split		Split ratio used
Cross-validation strategy		Details (e.g., k-fold, stratified)
Hyperparameters identified		Parameters selected for tuning
Hyperparameter tuning method		Method used (e.g., grid search, random search)
Training executed		Record any issues orobservations during training
Evaluation metrics defined		List of metrics (e.g., accuracy, precision, recall, F1-score).
Model performance recorded		Validation results for each model

Appendix 3: MUT Research Authorization Letter

MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Director, Postgraduate Studies

E-mail: [illegible]
Website: [illegible]

P.O. Box 75-10200
Murang'a, Kenya

Ref: MUT/ARP/PGS/19/2020 /VOL.1 **Date:** 10th July, 2024

Mr. DAVID KAIMENYI MARANGU- SC500/5100/2017

School of Computing and Information Technology,
Murang'a University of Technology,
P.O. Box 75-10,200
Murang'a.

Dear Mr. Marangu,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 1st July 2024 considered and approved your Doctoral research proposal entitled ***"A deep Ensemble model for Anomaly detection in Imbalanced energy consumption data"*** and appointed the following as supervisors:

1. Dr. Stephen Njenga - Murang'a University of Technology
2. Dr. Rachael Ndung'u – Murang'a University of Technology

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports forms can be accessed in the Student Portal.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.

Your responsibilities as a student will include, among others;

i. Maintain regular consultation with your supervisor(s), at least once a month
ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;

iii. Ensure quality work all through;

iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.

v. Publish two articles from your research findings in a refereed journal prior to thesis examination.

For any further clarification please contact the undersigned.

Yours Sincerely,

Prof. Geoffrey Muchiri, PhD
Director, Postgraduate Studies

DIRECTORATE OF POSTGRADUATE STUDIES
MURANG'A UNIVERSITY OF TECHNOLOGY
10 JUL 2024
P. O. Box 75 - 10200, MURANG'A
Tel: 0771463515, Email: info@mut.ac.ke

Copy to:

1. Registrar – Academics and Students Affairs
2. Dean, School of Computing and Information Technology

Appendix 4: NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Ref No: 151065

Date of Issue: 08/November/2024

RESEARCH LICENSE



This is to Certify that Mr., David kaimenyi Maranga of Murang'a University of Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Muranga on the topic: A DEEP ENSEMBLE MODEL FOR ANOMALY DETECTION IN IMBALANCED ENERGY CONSUMPTION DATA for the period ending : 08/November/2025.

License No: NACOSTI/P/24/41085

151065

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix 5: Publications

The following papers have been published from this thesis.

[1] D. K. Marangu, S. T. Njenga, and R. N. Ndung'u, 'Systematic Review of Models Used to Handle Class Imbalance in Anomaly Detection for Energy Consumption', *IJAIA*, vol. 15, no. 3, pp. 41–52, May 2024, doi: 10.5121/ijaia.2024.15304.

[2] D. K. Marangu, S. T. Njenga, and R. N. Ndung'u, 'A Comparative Analysis of Class Imbalance Handling Techniques for Deep Models in the Detection of Anomalies in Energy Consumption', *IJAIA*, vol. 15, no. 6, pp. 13–29, Nov. 2024, doi: 10.5121/ijaia.2024.15602.