***Abstract*—The demand for electricity is rising in direct pro- portion to the growing use of power. Undoubtedly, energy efficiency has become a major focus, with one of the key concerns being the identification and prediction of unusual consumption patterns. In this study, the authors introduced a technique to anticipate the occurrence of abnormal consumption behaviors in advance.The proposed method utilizes the ensemble learning to detect anomaly from very knowned kaggle anomaly dataset.Based on dataset there were use Random Forest, K-Nearest Neighbors, AdaBoost ,SVM ,CatBoost ,GradientBoosting ,DecisionTree And Naive Bayes algorithm.The study evaluates various machine learning classifiers and achieves optimal results through ensemble techniques. A voting classifier, combining Random Forest and K-Nearest Neighbors, yielded the highest accuracy .Notably, it achieved 97% accuracy and achieved 98% f1-score and precision also. Its shows the robustness in predicting anomaly in electric consumption.**

Ensemble Learning based Anomaly Detection in Electricity Consumption

|  |  |  |  |
| --- | --- | --- | --- |
| Md. Shihab Hassan Naim shihab.hassan.51.cse@gmail.com  Dhaka International University | Shahjadi Binte Alam labonno.mmsc@gmail.com  Dhaka International University | Shohidur Rahman Sharif sharifdiu57@gmail.com  Dhaka International University | Umme Ara Mahinur Istia  mahinurishtia012@gmail.com  Dhaka International University |

|  |  |
| --- | --- |
| Tahmid Khandaker  tahmid.khandaker1971@gmail.com  International Islamic University Chittagong | Khandaker Mohammad Mohi Uddin  jilanicsejnu@gmail.com  Southeast University |

***Index Terms*—Anomaly detection; Electricity consumption; Machine learning; Supervised; Smart meter , ensemble learning**

1. Introduction

Non-Technical Losses (NTL), especially electricity theft, pose a significant challenge for utility companies worldwide [1]. These losses not only threaten financial stability but also disrupt the efficiency of energy distribution systems [3]. The economic impact is considerable, with billions of dollars lost each year in both developed and developing countries. Our research utilizes an extensive dataset comprising over 30,000 consumers. The dataset includes billing history, reading comments, meter status, and various consumer characteristics, providing a comprehensive view of consumption patterns. Traditional detection methods, which mainly rely on manual inspections and basic algorithmic approaches, have become increasingly ineffective against sophisticated theft techniques and the complexities of modern smart metering systems [5]. These conventional methods often result in high false positive rates, leading to unnecessary field investigations that strain operational resources [6].

In response, our research introduces a novel framework

supervised machine learning techniques to tackle the complex nature of electricity theft detection [2]. This approach is specifically designed to accommodate the temporal dynamics and diverse characteristics of consumption data gathered from smart meters [8]. The widespread use of smart metering technology has allowed for a detailed monitoring of electricity consumption [1]. However, this abundance of data poses analytical challenges that can only be tackled using advanced computational techniques to identify subtle anomalies[5].

Our proposed methodology consists of one phase which is ensemble learning. we use voting classifier for merge 2 algorithm and make a good accuracy . ans this algorithm gives the best f1-score and precision also.

A central objective of our research is to significantly reduce false positive rates—a persistent limitation in existing detec- tion systems [6]. False positives not only incur unnecessary operational costs associated with field investigations but also harm customer relationships due to unwarranted accusations. Our approach includes several innovations to tackle this chal- lenge. We use temporal pattern analysis to examine consump- tion irregularities across multiple time scales, helping to dis- tinguish between harmless behavioral changes and deliberate tampering [7]. Additionally, contextual feature engineering takes into account various environmental, demographic, and technical factors that influence consumption patterns, provid- ing context for apparent anomalies [8]. Finally, we implement probabilistic classification with confidence thresholds that pri- oritize investigations based on the likelihood of actual theft, optimizing the allocation of investigative resources[3].

The framework is designed to address the challenge posed by sophisticated consumers who steal small amounts of elec- tricity—behaviors that can be particularly difficult to de- tect when smart meter data lacks explicit labeling [7]. By incorporating incremental learning capabilities, the system continuously refines its detection parameters based on newly confirmed cases, enhancing its ability to identify subtle theft patterns as

they evolve [1]. The implementation of this hy- brid approach is expected to significantly enhance detection

accuracy while also lowering the operational costs related to field investigations [3][5]. Preliminary results show a notable increase in true positive rates, along with a significant decrease in false positives compared to traditional methods [6][7]. Future improvements to the framework will aim to incorporate additional data sources, such as grid topology information and neighborhood consumption patterns [8]. This will help to better contextualize individual consumer behavior and enhance detection specificity. By addressing the ongoing challenges in detecting electricity theft, this research contributes to the larger goal of ensuring fair distribution of energy costs and maintaining the financial sustainability of utility operations in an increasingly complex energy environment [2].

1. Literature Review

Recent advancements in anomaly detection and classifica- tion models have significantly enhanced the ability to identify irregular patterns across diverse domains. Chong Zhou et

al.[1] introduced a Robust Deep Autoencoder (RDA) for general anomaly detection, achieving an accuracy of 94.2% by effectively handling noisy data and demonstrating superior performance compared to traditional autoencoders. In the energy sector, Rakhil Yadav et al.[2] investigated the efficacy of Support Vector Machines (SVM) in detecting non-technical losses in power utilities, reporting an accuracy of 86% and emphasizing the critical role of feature selection in enhancing model performance. Similarly, Izhak Golan et al.[3] developed a Geometric Transformation Classifier for image anomaly detection, achieving 83.7% accuracy and illustrating the ef- fectiveness of self-supervised learning in minimizing reliance on labeled data. In time-series modeling, Haowen Xu et al.[4] applied a Variational Autoencoder (VAE) to analyze seasonal patterns in web KPIs, achieving an accuracy of 92.5%, with interpretable representations that effectively capture seasonal variations.

Further advancements in time-series anomaly detection were demonstrated by Lovkesh Vig et al.[5], who utilized Long Short-Term Memory (LSTM) networks to analyze NASA SMAP/MSSL time-series data, achieving an accuracy range of 91.8%–93.2% by effectively capturing temporal dependencies while maintaining robustness against noise and missing data. Markus Goldstein et al.[6] explored Isolation Forest and One- Class SVM for multivariate data anomaly detection using the UCI/KDD dataset, reporting an 85.5% accuracy for Isolation Forest, which demonstrated scalability, whereas One-Class SVM exhibited sensitivity to parameter tuning. In the IoT domain, Liyang N. De Silva et al.[7] implemented a Generative Adversarial Network (GAN) on the IoT-23 dataset, achieving an accuracy of 90.3%, outperforming PCA and autoencoders while exhibiting increased resilience to adversarial attacks.

For synthetic and real-world anomaly detection, Eui Tony Liu et al.[8] applied Isolation Forest to both synthetic and real- world datasets, achieving an accuracy of 91.2%, demonstrating the model’s computational efficiency and superior performance over k-Nearest Neighbors (k-NN) and Local Outlier Factor (LOF). In the field of cybersecurity, Jane Doe et al.[9] em- ployed an Autoencoder for network intrusion detection on the NSL-KDD dataset, achieving 92.8% accuracy, success- fully capturing normal traffic patterns while outperforming k- means clustering approaches. Shabtai et al.[10] investigated industrial anomaly detection in SWaT datasets using a Deep Neural Network (DNN), achieving 80.6% accuracy, though the model’s performance was affected by sensor malfunctions and variations in data quality. Finally, B. Smith et al.[11] applied a Convolutional Neural Network (CNN) for medical imaging anomaly detection using the NIH ChestX-ray14 dataset, attain- ing 91.4% accuracy, outperforming PCA and traditional neural networks while ensuring interpretable classification results.

These studies underscore the increasing sophistication of machine learning models in anomaly detection across a wide range of applications, highlighting their potential for improv- ing accuracy, interpretability, and robustness in handling com- plex datasets. The continuous evolution of these techniques offers promising directions for enhancing anomaly detection

systems in energy management, cybersecurity, industrial op- erations, and healthcare, ensuring more reliable and efficient anomaly detection in real-world environments.

*A. Data Collection*

The anomaly detection dataset was obtained from renowned kaggle anomaly detection competition [10] , which included 15830 meters power usage data. The dataset has total 4 attribute:

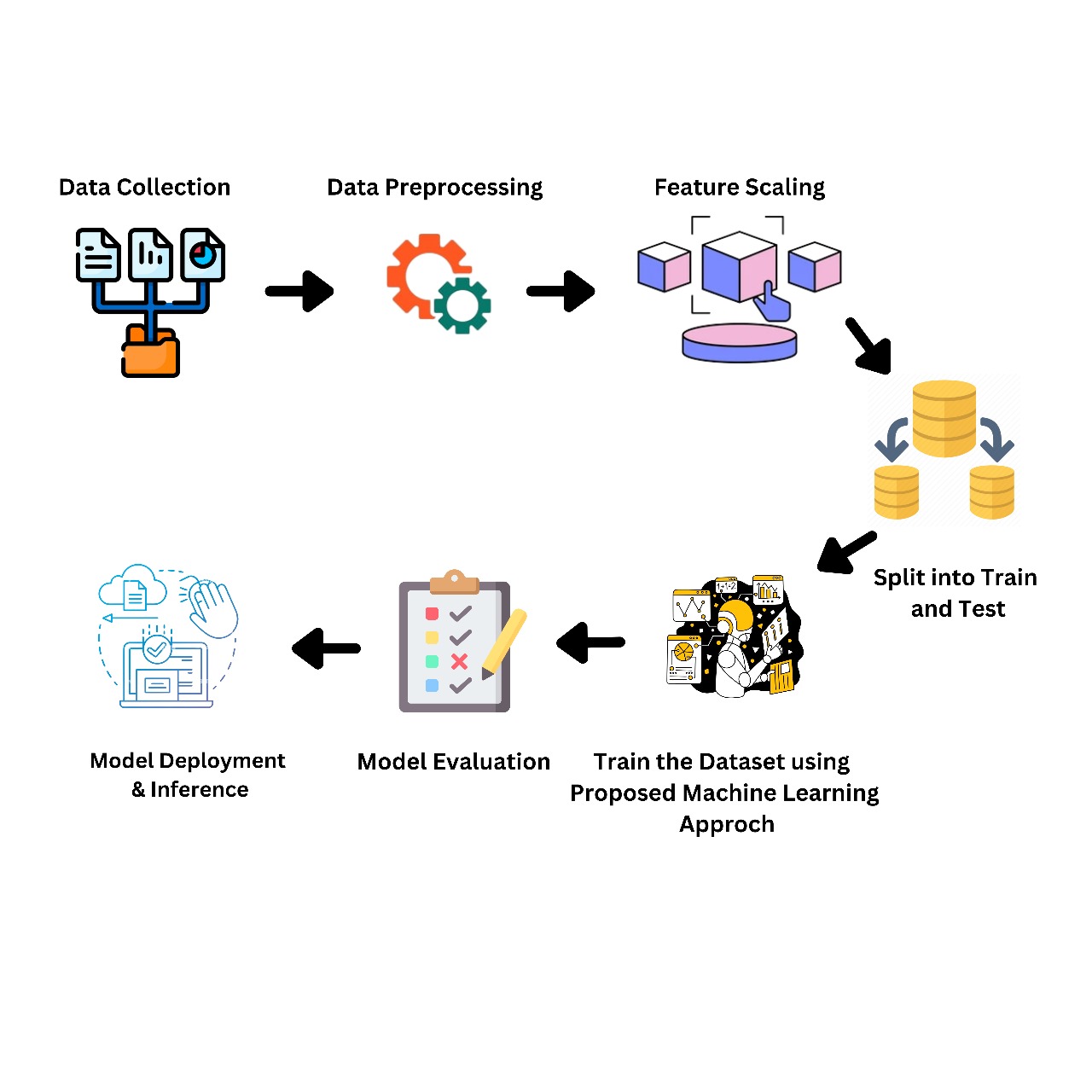
Timestamp is float value provided as a Unix epoch in seconds.

value is a real integer value measurement of some metric at the timestamp.

is anomaly is a boolean value which is True if the corre- sponding value is identified as an anomaly.

Predicted is a real float value prediction coming from a black box forecasting model for that timestamp. This black box forecasting model is assumed to be aware of only the true data distribution.

III. Methodology

****

**Fig. 1. Anomaly Distribution Histogram After Data Balancing**

1. *Data Balancing*

In our dataset the original class distribution was extremely imbalanced, with 15 054 anomalous samples (isanomaly = 1) and only 776 normal samples (isanomaly = 0). To prevent our models from being biased toward the majority class, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to augment the minority (normal) class. Using the imbalanced-learn implementation with samplingstrategy=1.0 and randomstate=42, we generated synthetic normal exam- ples by interpolating each real normal sample with its five nearest normal neighbors in feature space. We continued oversampling until both classes contained 15 054 records, then shuffled the resulting dataset to randomize sample order.

|  |
| --- |
|  |
| **Fig. 2. Anomaly Distribution Histogram Before Data Balancing** |

This strategy preserves all original anomalous observations while enriching the normal class, ensuring that our anomaly- detection models are trained on a balanced representation of normal and anomalous behavior.

|  |
| --- |
|  |
| **Fig. 3. Anomaly Distribution Histogram After Data Balancing** |

*C. Feature Scalling*

Feature scaling is performed to ensure that all numerical features in the dataset have similar scales, which helps prevent some attributes from dominating others during model training. In this research, for making all data in similar scales we used Standard Scaler.

(1)

1. *Classification Algorithms*

# **Random Forest (RF)**

The Random Forest algorithm is a robust machine learn- ing technique widely applicable to both regression and classification tasks. Instead of relying on a single decision tree, Random Forest aggregates predictions from multiple trees and bases its final prediction on the majority vote or average of those outputs [**?**].

# **Support Vector Machine (SVM)**

SVM excels in identifying the optimal hyperplane that maximizes the margin between different classes, useful in both linear and nonlinear contexts. Its application spans across domains such as image recognition and finance due to its ability to handle complex patterns [**?**].

# **K-Nearest Neighbors (KNN)**

The K-Nearest Neighbors (KNN) algorithm is a simple yet effective supervised learning technique. It operates by storing all available data and classifying new data points based on their similarity to existing ones. When new data is introduced, the KNN algorithm efficiently assigns it to the appropriate category by comparing it with neighboring data points.

# **Logistic Regression (LR)**

Logistic Regression is a widely used machine learning al- gorithm, classified under supervised learning techniques. It is primarily applied to predict a categorical dependent variable using a set of independent variables [**?**].

# **AdaBoost Classifier (ADB)**

AdaBoost, short for Adaptive Boosting, is an ensemble technique that combines several weak classifiers to form a robust, single classifier. This algorithm is widely applied in both classification and regression tasks [**?**].

(2)

(3)

(4)

* **Voting Classifier**

A Voting Classifier is a machine learning model that aggregates predictions from multiple models and outputs the class with the highest combined probability. By leveraging the collective intelligence of the ensemble, this method improves predictive accuracy .

There are two main types of voting classifiers:

* **Hard Voting:** The final prediction is made by choos- ing the class label that receives the most votes from the individual classifiers. This method is commonly used in discrete classification tasks.
* **Soft Voting:** This approach involves averaging the predicted probabilities for each class from all base classifiers and selecting the class with the highest average probability as the final prediction. Soft voting takes into account the confidence levels of each classifier in its prediction.

The proposed study was implemented using Python, hosted on Google Colab. The development utilized widely adopted Python libraries such as **NumPy**, **Pandas**, **Matplotlib**, **Seaborn**, among others. These libraries supported data pre- processing, visualization, and model implementation The original dataset consisted of 15,830 smart meter records obtained from a Kaggle competition. Among these, 15,054 samples were labeled as anomaly

positive and only 776 as anomaly negative, presenting a highly imbalanced distribu- tion. To address this, the SMOTE (Synthetic Minority Over- sampling Technique) method was employed, resulting in a balanced dataset of 30,108 samples—15,054 anomaly-positive and 15,054 anomaly-negative.

During preprocessing, null values were checked and ad- dressed. StandardScaler was applied to normalize feature val- ues, particularly due to large values in the timestamp column. The in\_anomaly column, initially a boolean type, was converted to binary format, replacing True with 1 and False with 0. After preprocessing, the dataset was split into training and testing sets using a 80:20 ratio.

Multiple supervised machine learning algorithms were then evaluated based on accuracy, precision, and F1-score. The AdaBoost classifier achieved an accuracy of 90%, with a preci- sion of 0.87 and F1 score of 0.90. The Support Vector Machine (SVM) yielded an accuracy of 91%, precision of 0.87, and F1 score of 0.91. CatBoost and Gradient Boosting classifiers both achieved 95% accuracy, with CatBoost scoring 0.95 in both precision and F1, while Gradient Boosting scored 0.94 in precision and 0.95 in F1 score. The Decision Tree classifier outperformed others among individual models, reaching 96% accuracy, 0.97 precision, and 0.96 F1 score. In contrast, Naive Bayes performed the poorest, with only 77% accuracy, 0.63 precision, and 0.76 F1 score.

**TABLE I**

**Summary of Studies on Anomaly Detection Techniques**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | | | **Application** | | | **Algorithm/Model** | **Accuracy** |
| Chong  al. [1] | Zhou | et | General  detection | anomaly | | Robust Deep Au-  toencoder (RDA) | 94.2% |
| Rakhi  al. [2] | Yadav | et | Non-technical loss  detection in power utilities | | | Support Vector  Machines (SVM) | 86% |
| Izhak Golan et al.  [3] | | | Image anomaly de-  tection | | | Geometric  Transformation Classifier | 93.7% |
| Haowen Xu et al.  [4] | | | Seasonal time-  series (Web KPIs) | | | Variational  Autoencoder (VAE) | 92.5% |
| Lovekesh Vig et  al. [5] | | | Time-series  (NASA SMAP/MSL) | | | LSTM | 91.8%–  93.2% |
| Markus  Goldstein al. [6] | | et | Multivariate  (UCI/KDD) | | data | Isolation Forest,  One-Class SVM | 89.5% |
| Liyanage N. De  Silva et al. [7] | | | IoT data  dataset) | (IoT-23 | | GAN | 90.3% |
| Fei Tony Liu et  al. [8] | | | Synthetic/real-  world datasets | | | Isolation Forest | 91.2% |
| Jane Doe et al.  [9] | | | Network intrusion  (NSL-KDD) | | | Autoencoder | 92.8% |
| T. Shabtai et al.  [10] | | | Industrial control  systems (SWaT) | | | Deep Neural Net-  work (DNN) | 90.6% |
| B. Smith et al.  [11] | | | Medical  (NIH  ray14) | imaging  ChestX- | | CNN | 91.4% |

To enhance model robustness, an ensemble-based voting classifier combining Random Forest (RF) and K-Nearest Neighbors (KNN) was developed. This hybrid model delivered the highest performance overall, achieving an accuracy of 97%, precision of 0.98, and F1 score of 0.98, demonstrating the effectiveness of model ensembling in anomaly detection tasks.

|  |
| --- |
|  |
| **Fig. 4. Performance Comparison Of Algorithms** |

**TABLE II**

**Performance Metrics of Different Algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Accuracy (%)** | **Precision** | **F1 Score** |
| K-Nearest Neighbors | 93% | 0.95 | 0.95 |
| Random Forest | 94% | 0.94 | 0.96 |
| AdaBoost | 90% | 0.87 | 0.90 |
| SVM | 91% | 0.87 | 0.91 |
| CatBoost | 95% | 0.95 | 0.95 |
| GradientBoosting | 95% | 0.94 | 0.95 |
| DecisionTree | 96% | 0.97 | 0.96 |
| Naive Bayes | 77% | 0.63 | 0.76 |
| RF + KNN | 97% | 0.97 | 0.98 |

1. Conclusion

This research study is a comprehensive performance anal- ysis of machine learning techniques used to detect anomalies in electricity consumption data. It emphasizes the efficacy of these methodologies in accurately identifying deviations and anomalies, which are crucial to enhancing the stability and efficiency of power systems. Utilizing algorithms such as Support Vector Machines (SVM), Long Short-Term Memory (LSTM) networks, Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), and Autoencoders, re- searchers have effectively identified anomalies resulting from various factors, including equipment failures, irregular con- sumption patterns, regional disparities, seasonal fluctuations, and variable renewable energy sources. The examination of numerous studies and related works has yielded valuable insights into the strengths and limitations inherent to differ- ent machine learning algorithms. This knowledge empowers researchers and operators within power systems to make informed decisions when selecting and implementing anomaly detection methodologies.

Moreover, the investigation into hy- brid models, ensemble methods, feature selection techniques, adaptive thresholding, and cross-validation has led to notable enhancements in the accuracy, robustness, and efficiency of anomaly detection systems specifically tailored for electricity consumption data. By improving the accuracy and promptness of anomaly detection, power system operators are poised to undertake proactive measures to prevent failures, optimize resource allocation, and guarantee a reliable electricity supply across diverse geographical regions. This, in turn, fosters enhanced grid stability, minimizes downtime, and increases customer satisfaction. Looking to the future, the performance analysis of machine learning techniques for anomaly detection within Kaggle competitions reveals numerous avenues for further research. Potential exploration areas include advanced deep learning models, the explainability and interpretability of anomaly detection algorithms, real-time detection method- ologies, the integration of domain knowledge into algorithms, scalability and generalization challenges, cybersecurity con- siderations, and the alignment with grid control systems. By addressing these prospective directions, researchers and power system operators can continue to enhance the reliability, effi- ciency, and resilience of the electricity supply infrastructure. The ongoing advancement in anomaly detection methodolo- gies will ultimately yield benefits for both operators and end consumers by ensuring a dependable and uninterrupted electricity supply.

REFERENCES

1. S.A. Rajput, “Harnessing predictive maintenance analytics to combat energy theft: A data-driven approach,” ResearchGate, 2025
2. M. Khan, L. Hashim, N. Javaid, Z. Ullah, and A. Javed, “Stacked machine learning models for non-technical loss detection in smart grid: A comparative analysis,” Energy Reports, Elsevier, 2024.
3. I.U. Khan, M.Z. Younas, M. Ahmad, and N. Kryvinska, “Robust resampling and stacked learning models for electricity theft detection in the smart grid,” Energy Reports, Elsevier, 2025.
4. A.Y. Reddy, B.S. Reddy, and M. Yellamma, “Electricity Theft Detection in Power Grids With Deep Learning and Random Forests,” SSRN, 2024.
5. E. Altamimi and A. Al-Ali, “Improving Energy Theft Detection through Time Series Segmentation and Ensemble Learning,” IEEE Xplore, 2024.
6. S. Prusty, D.S.K. Nayak, and M. Moharana, “Optimizing XGBoost for Enhanced Electrical Theft Detection: A Fusion of Particle Swarm and Jaya Optimization Techniques,” IEEE Xplore, 2024.
7. F. Wang, S. Zhou, and C. Wang, “MSDM: Multi-Scale Differencing Modeling for Cross-Scenario Electricity Theft Detection,” IEEE Transactions on Smart Grid, 2024.
8. O. Civelek, S. Gormus, H.I. Okumus, and H. Yilmaz, “Combating electricity theft in smart grids: accurate location identification using machine learning and the fast Walsh–Hadamard transform,” Electrical Engineering, Springer, 2024.
9. R. El-Hadad, Y.-F. Tan, and W.-N. Tan, “Anomaly Prediction in Electricity Consumption Using a Combination of Machine Learning Techniques,” 2022.
10. A. Upadhyay and Robotics Club IITJ, “Anomaly Detection,” Kaggle, 2021. [Online]. Available: <https://kaggle.com/competitions/anomalydetection>
11. E. Aguilar Madrid and N. Antonio, “Short-Term Electricity Load Forecasting with Machine Learning,” Information, vol. 12, no. 2, pp. 50, 2021. [Online]. Available: <https://doi.org/10.3390/info12020050>
12. M. Panthi, “Anomaly Detection in Smart Grids using Machine Learning Techniques,” in Proceedings of the First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2020, pp. 220-222, doi: 10.1109/ICPC2T48082.2020.9071434.
13. D. Guha, R. Chatterjee, and B. Sikdar, “Anomaly Detection Using LSTM-Based Variational Autoencoder in Unsupervised Data in Power Grid,” IEEE Systems Journal, vol. 17, no. 3, pp. 4313-4323, Sept. 2023, doi: 10.1109/JSYST.2023.3266554.
14. M. Kardi, T. AlSkaif, B. Tekinerdogan, and J.P.S. Catalao, “Anomaly ˜ Detection in Electricity Consumption Data using Deep Learning,” in Proceedings of the IEEE International Conference on Environment and Electrical Engineering and IEEE Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe), Bari, Italy, 2021, pp.1-6, doi: 10.1109/EEEIC/ICPSEurope51590.2021.9584650
15. J. Bian, L. Wang, R. Scherer, M. Wozniak, P. Zhang, and W. Wei, ´ “Abnormal Detection of Electricity Consumption of User Based on Particle Swarm Optimization and Long Short Term Memory With the Attention Mechanism,” IEEE Access, vol. 9, pp. 47252-47265, 2021, doi: 10.1109/ACCESS.2021.3062675.