

Intrusion Detection in Carbon Neutral Data Centers Based on Integrated Renewable Energy Generation

Min ZHANG* 

*Cangzhou Normal University, Cangzhou 061001, China

(czsf031038@sina.com)

Received: 27.06.2026 Accepted: 14.09.2026

Abstract- Energy systems play a significant role in the current world, involving not only fossil fuels but also renewable energy sources such as solar and wind, as well as battery storage. The production of renewable energy is crucial for reducing carbon emissions worldwide and supporting the development of carbon-neutral infrastructure, such as data centers, that involve energy production alongside cybersecurity concerns. Therefore, predicting the intrusion status of interconnected data centers powered by renewable energy is crucial to ensure their continued performance. Machine learning (ML) is effective in developing precise models of nonlinear relationships among variables in energy production and cybersecurity. To ensure methodological robustness, multicollinearity among predictor variables is assessed using the Variance Inflation Factor method, with a threshold of 5. In addition, to ensure representational fairness between classes in imbalanced datasets, SMOTE-ENC is used for oversampling. Moreover, to ensure the model's reliability, K-fold cross-validation and the Brier score are implemented. Three base classifiers are used, optimized using the Heavy Rain Optimization Algorithm (HROA) and the Kangaroo Optimization Algorithm (KOA). Additionally, to compare model performance, a Wilcoxon signed-rank test is used, and Copula analysis is used to assess interdependencies among variables. In summary, the results show that the HGBC (H1) performs better in terms of accuracy and reliability, with the lowest Brier score.

Keywords Renewable Energy, Intrusion Status, Energy Sources, Copula, Brier Score.

1. Introduction

Energy systems form part of complex infrastructures that have the main functions of generating, transmitting, distributing, and managing energy to accommodate the wide range of needs of modern society [1], [2], [3]. For several decades, these infrastructures have primarily relied on fossil fuels [4], including coal, oil, and gas, which have been key drivers of development and innovation while also contributing to environmental degradation through greenhouse gas emissions [5]. The realization of the consequences of global warming, coupled with the alarming rise in global energy demand driven by population growth, has underscored the importance of transitioning from conventional, unsustainable energy systems to sustainable ones to ensure long-term environmental sustainability [6].

Renewable energy systems are an essential part of the ongoing transition, as their use helps reduce dependency on fossil fuels and carbon emissions [7], [8], [9]. The use of solar photovoltaic cells [10], wind turbines [11], [12], [13], hydroelectric power plants [14], and bio-energy plants helps to reduce the carbon footprint [15], [16]. Achieving carbon

neutrality, a state characterized by balancing carbon dioxide emissions with offsets to reach zero carbon emissions, can be achieved by integrating renewable energy into existing infrastructure. The integration of renewable energy with technologies such as energy storage and power management enhances its flexibility and reliability and hastens the transition to sustainable energy infrastructure [17], [18].

Despite these environmental advantages, the modernization and digitalization of the energy system have also introduced new cybersecurity risks, particularly due to the susceptibility of smart grid systems and monitoring devices to cyber threats and attacks [19]. Energy systems are more likely to be attacked because they depend on communication and data control systems used to operate and control them. In this regard, predicting intrusion status has become vital for maintaining the sustainability of these systems, protecting the information used within them, and ensuring that the energy supply is not disrupted [20], [21].

Therefore, the use of Machine Learning (ML) can offer a strong solution for predicting intrusion status, owing to its ability to analyze massive, complex datasets from sensors,

communication records, and control systems to identify hidden patterns associated with cyber intrusions [22]. By integrating ML, it is possible to develop intelligent intrusion detection systems that respond to cyber intrusions in real time [23], [24]. The use of data-driven models can improve estimate accuracy through iterative training and validation, making them suitable for complex, changing energy systems that incorporate renewable energy. There is a substantial literature on this topic that points to its significance and potential [25].

Alimi et al. conducted an empirical evaluation of ML to prevent intrusions in power system networks. The effectiveness, accuracy, and applicability of K-NN, decision trees, Naive Bayes, random forests, and AdaBoost classifiers were evaluated for predicting and detecting cyber intrusions in smart grid systems [26]. Gokulraj and Venkatramanan proposed a sophisticated ML framework to improve security and detect anomalies in cyber-physical microgrids with inverters, focusing on cyber and operational issues. They assessed intelligent models to improve the accuracy, reliability, and protection of distributed energy systems [27].

Hashmi et al. also developed an ML-based architecture for fault mitigation in renewable energy systems, including an evaluation of economic performance. In this regard, the study used predictive models to diagnose faults in renewable energy systems while assessing their cost-effectiveness and operational benefits to improve sustainability [28]. In another study, Aygul et al. conducted a benchmark evaluation of ML techniques for predicting transient stability in power grids with high penetration of renewable energy sources and exposure to cyber-attack scenarios. The authors tested and compared the models' accuracy for robustness and reliability [29].

The novelty of this study lies in the development of a comprehensive and robust machine learning framework for intrusion prediction in renewable energy-powered data centers by integrating advanced preprocessing, hyperparameter optimization, statistical validation, calibration, and dependency analysis within a unified workflow. Three complementary classifiers were selected to represent different learning paradigms: Linear Discriminant Analysis (LDA) as an efficient linear classifier, Extra Trees Classifier (ETC) as a randomized ensemble model for capturing complex nonlinear relationships, and Histogram Gradient Boosting Classifier (HGBC) as a gradient boosting technique recognized for its high predictive performance and computational efficiency on structured tabular data.

To optimize model hyperparameters, the Heavy Rain Optimization Algorithm (HROA) and the Kangaroo Optimization Algorithm (KOA) were adopted because they provide population-based global search strategies with balanced exploration and exploitation, making them suitable for the complex, multimodal search spaces associated with machine learning hyperparameter optimization. In contrast to Bayesian Optimization, which is highly effective for low-dimensional and computationally expensive optimization problems but relies on surrogate modeling that may become less efficient as the search space grows, HROA and KOA

directly explore the solution space without requiring probabilistic surrogate models. Similarly, unlike Genetic Algorithms, which rely on crossover and mutation operators and may experience premature convergence or reduced population diversity, HROA and KOA employ adaptive search mechanisms that promote broader exploration while maintaining competitive convergence behavior. These characteristics motivated their selection for optimizing the proposed classifiers.

Furthermore, the proposed framework integrates Variance Inflation Factor (VIF) analysis to mitigate multicollinearity, SMOTE-ENC to address class imbalance, the Brier Score to assess model calibration, the Wilcoxon signed-rank test to verify the statistical significance of performance differences, and Copula analysis to investigate dependency structures among renewable energy and cybersecurity variables. Unlike conventional intrusion detection studies that primarily focus on predictive accuracy, the proposed framework combines advanced optimization with statistical validation, reliability assessment, and dependency analysis, providing a more rigorous and practically applicable methodology for intrusion prediction in renewable energy-powered data center environments.

2. Mathematical Methods

2.1. Proposed models

➤ Linear Discriminant Analysis (LDA)

LDA is a supervised learning method that is both a statistical technique and an ML algorithm, applicable to both classification and dimensionality reduction problems. The method works by finding the linear combination of features that best discriminates between two or more classes of objects. Unlike unsupervised learning approaches such as Principal Component Analysis, which seek to maximize variance, LDA seeks to maximize the ratio of between-class to within-class variance, meaning that the projected data points are as far apart as possible between classes while being as close as possible within classes. It applies to areas such as pattern recognition, biology, and finance, among others, and is computationally efficient because it produces linear decision boundaries [30].

➤ Extra Trees Classifier (ETC)

ETC [31], similar to Random Forest, is an ensemble learning model used for classification that creates numerous decision trees and combines their predictions. In contrast to the Random Forest algorithm, Extra Trees uses more randomness by randomly selecting thresholds for splitting rather than choosing the best split for each feature and often uses all the data rather than samples. This increased randomness results in lower variance, reducing overfitting and improving performance. The prediction is made by averaging all the probabilities or by voting across all the trees. Extra Trees is computationally efficient, particularly for high-dimensional data, and can be used for complex nonlinear classification problems.

➤ Histogram-Based Gradient Boosting Classification (HGBC)

HGBC [32] is an emerging ML algorithm with promising potential to boost the computational speed of traditional gradient boosting algorithms. The traditional gradient boosting algorithm relies on building weak models, typically decision trees. These decision trees are constructed sequentially, and each attempts to correct the mistakes of the previous one. Although this approach has been successful in practice, it is computationally expensive. The HGBC algorithm addresses this problem by transforming continuous features into a histogram-based feature set. This approach minimizes the amount of data to be processed.

Another important advantage of HGBC over traditional gradient boosting methods is its scalability. The use of histogram binning and parallel processing reduces the required memory and time. Despite these advantages of HGBC, gradient boosting maintains strong predictive performance. As such, it becomes useful in tasks such as image recognition and fraud detection. However, as with other boosting methods, proper hyperparameter tuning is required to avoid overfitting.

2.2. Metaheuristic Algorithms

➤ Heavy Rain Optimization Algorithm (HROA)

HROA is a nature-inspired metaheuristic optimization algorithm that mimics the nature of heavy rain and the natural flow of water. The algorithm simulates the process of raindrops falling, accumulating, joining, and flowing downward. In other words, it simulates raindrops passing through the solution space to reach the optimal solution. In HROA, raindrops move through the solution space as they pool, run off, and intensify flow, maintaining a balance between exploration and exploitation. Weak solutions are modified or guided to more suitable regions of the solution space, enabling the algorithm to reach the global optimum. Like other population-based optimization algorithms, such as Particle Swarm Optimization, HROA can be used to solve complex, nonlinear, and high-dimensional optimization problems [32].

➤ Kangaroo Optimization Algorithm (KOA)

Based on kangaroo locomotion, KOA uses both local and global search processes in a metaheuristic search. This metaheuristic search uses two population groups: mother kangaroos, who explore the solution space, and offspring, who align with their mother to examine their region. More experienced youngsters use both local and global search processes to update their current positions and build complicated solutions in the solution space. Mother kangaroos undertake exploratory moves to locate the ideal areas during this time.

The strategy explores the solution space using a combination of general and specialized search strategies, imitating the mechanisms used by kangaroos and other living things. Convergence avoidance, a significant optimization problem, is influenced by the natural bond between a mother

and her joey. As a result, this approach avoids local optimal traps and allows for a more thorough search for optimal solutions. KOA's hybrid form potentially allows it to handle a wide range of optimization challenges [33].

2.3. Additional Methods

The Variance Inflation Factor (VIF) is a measure of multicollinearity in the data used in regression analysis. The VIF measures the extent to which the variance of a predictor is increased by its relationship with other predictors. A threshold of 5 indicates the potential for problematic multicollinearity, and the variable may need to be removed to improve the model's reliability [34]. K-fold cross-validation divides the dataset into k equal halves, trains the model on $k-1$ folds, and then repeatedly tests on the remaining fold to guarantee robust performance evaluation and reduce bias from a single train-test split [35]. The model's performance is assessed using the Wilcoxon Signed-Rank Test, a nonparametric test that compares paired observations and determines whether the median of the differences is substantially different from zero [36], [37].

Furthermore, SMOTE-ENC [38] generates synthetic samples from the minority class while preserving the structure of categorical variables, addressing class imbalance in datasets with a combination of numerical and categorical features. Sensitivity analysis using Copula is a method for measuring the extent to which dependent input variables affect model results by separately analyzing their interdependencies, independent of their marginal distributions. Copulas are useful in analyzing nonlinear relationships and are employed to determine the degree to which associated uncertainties affect the system's performance [39], [40].

2.4. Performance Evaluators

Several evaluation metrics are used to assess the efficiency of ML algorithms, especially for error management, since using a single metric may not accurately reflect overall model performance. The Matthews Correlation Coefficient (MCC) is one of the most comprehensive evaluation metrics, as it accounts for true positives, true negatives, false positives, and false negatives, providing an overall evaluation of the model, especially when the dataset is imbalanced. The Area Under the Receiver Operating Characteristic Curve (AUC) is commonly used to evaluate the discriminative power of models across different classes in a dataset.

In cases of class imbalance, Accuracy can be misleading because the model can achieve high Accuracy by simply predicting the majority class, resulting in high Accuracy while performing poorly on the minority class. Precision is defined as the ratio of true positives to the total number of predicted positive classes, while Recall is defined as the ratio of true positives to the total number of actual positive classes. The F1-score is the harmonic mean of Precision and Recall; it is particularly useful in cases of class imbalance because it balances the two metrics. In addition, per-class error assessment provides a more comprehensive evaluation by determining the error for each class.

Markedness evaluates the model's ability to differentiate between positive and negative examples by comparing its Accuracy and false-positive rate. In cases where the model provides probabilistic output, especially for binary classification problems where the output can be 0 or 1, the Brier Score plays a vital role in assessing the model's performance, as it measures the mean squared error between the output probabilities and the actual outputs. A Brier Score of 0 signifies a perfect output, while a higher score signifies a lower output, given that the Brier Score essentially calculates the mean squared error of the output [34].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$MCC = \frac{(TP * TN) - (FP + FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

$$Class_wise\ error = \frac{FP + FN}{TP + TN + FP + FN} \quad (6)$$

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (7)$$

$$Markedness = Precision + \frac{TN}{TN + FN} - 1 \quad (8)$$

$$Brier\ Score = \frac{1}{N} \sum_{i=1}^N (P_i - O_i)^2 \quad (9)$$

3. Data Collection

The dataset employed in this study is the publicly available Renewable Cyber Defense Dataset obtained from the Kaggle repository [41]. It is a synthetic benchmark dataset designed to emulate the operational characteristics of

renewable energy-powered carbon-neutral data centers by integrating renewable energy management variables with cybersecurity event logs. The dataset comprises 1,500 samples with 12 input variables and one target variable (Intrusion Status). The predictor variables describe energy production and consumption, carbon emissions, load balancing, energy scheduling optimization, network traffic, access control, blockchain verification, anomaly scores, and intrusion attempt frequency, providing a representative benchmark for investigating the relationship between energy management and cybersecurity. Although the dataset is organized as time-series records with timestamps to emulate continuous monitoring, it is intended as a benchmark dataset for developing and evaluating intrusion prediction models rather than representing measurements collected from a real-world operational data center. Its public availability and comprehensive integration of energy and cybersecurity attributes make it a suitable benchmark for assessing machine learning methods in renewable energy-powered data center environments.

Table 1 shows that although the original and balanced data have the same minimum and maximum values for battery, grid, solar, and wind sources, significant differences emerge in the variability and security metrics. The original data's energy generation shows a higher standard deviation, whereas the balanced data's mean increases slightly, with a reduced standard deviation of approximately 514. The same phenomenon holds for the Energy Consumed attribute, where the original data's mean of 977 and standard deviation of 479 become slightly more consistent in the balanced data, with a reduced variability of 466.

The original data's carbon emissions reveal a slight reduction in the mean, from approximately 81 to 79.4, with reduced variability. Energy Scheduling remains constant across both data sets, but for balanced data, it has a slightly higher average (0.798 and 0.795) and lower variability, suggesting better optimization. Network Traffic Volume ranges between 10,550 and 10,570 in both datasets but exhibits significantly lower variability in the balanced set. The average System Anomaly Score is around 0.50 in both datasets, with lower variability in the balanced set. Additionally, the significant variation is in Intrusion Attempts: in the original data, it averages 4.56 (median 5), but drops significantly to 3.28 (median 2) in the balanced data.

Table 1. Comparative Descriptive Statistics of Original and Balanced Datasets Across Energy Sources

Variables	Energy Source: Battery						
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1091.0	522.0	1085.6	636.0	1543.5
Energy_Consumed	152.1	1799.6	976.8	478.6	971.1	563.0	1393.4
Carbon_Emissions	10.05	199.7	81.03	56.12	57.91	36.81	126.0
Energy_Scheduling	0.6	0.99	0.795	0.112	0.79	0.7	0.89
Network_Traffic_Volume	1011	19978	10552.5	5462.3	10320	5907	15310.5

System_Anomaly_Score	0	1	0.502	0.287	0.520	0.260	0.750
Intrusion_Attempt	0	9	4.563	2.870	5	2	7
Energy Source: Grid							
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1090.6	522.4	1085.6	635.5	1543.5
Energy_Consumed	152.1	1799.6	977.0	478.5	971.12	565.1	1393.4
Carbon_Emissions	10.1	199.7	81.05	56.10	57.93	36.8	126.0
Energy_Scheduling	0.600	0.990	0.795	0.112	0.790	0.700	0.890
Network_Traffic_Volume	1011	19978	10544.5	5464.2	10304	5894.5	15290
System_Anomaly_Score	0	1	0.502	0.287	0.510	0.255	0.750
Intrusion_Attempt	0	9	4.561	2.870	5	2	7
Energy Source: Solar							
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1091.4	522.2	1086.6	636.2	1544.2
Energy_Consumed	152.1	1799.6	977.3	479.0	971.5	563.3	1395.1
Carbon_Emissions	10.1	199.7	80.99	56.12	57.90	36.78	126.0
Energy_Scheduling	0.600	0.990	0.795	0.112	0.790	0.700	0.890
Network_Traffic_Volume	1011	19978	10557.3	5463.7	10320.5	5911	15315
System_Anomaly_Score	0	1	0.502	0.287	0.515	0.258	0.750
Intrusion_Attempt	0	9	4.561	2.870	5	2	7
Energy Source: Wind							
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1090.6	522.0	1085.2	635.9	1542.6
Energy_Consumed	152.1	1799.6	976.4	478.6	969.7	562.6	1392.3
Carbon_Emissions	10.05	199.7	80.96	56.08	57.9	36.79	126.0
Energy_Scheduling	0.600	0.990	0.795	0.112	0.790	0.700	0.890
Network_Traffic_Volume	1011	19978	10547.3	5460.5	10312	5903	15298
System_Anomaly_Score	0	1	0.502	0.287	0.515	0.260	0.750
Intrusion_Attempt	0	9	4.562	2.871	5	2	7
Balanced							
Energy Source: Battery							
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1095.9	514.2	1102.9	648.4	1544.7
Energy_Consumed	152.1	1799.6	978.7	465.9	960.3	589.7	1383.3
Carbon_Emissions	10.05	199.7	79.37	51.74	59.25	38.57	117.3

Energy_Scheduling	0.600	0.990	0.798	0.104	0.795	0.714	0.884
Network_Traffic_Volume	1011	19978	10569.5	5461.2	10345.5	6122.5	15209.3
System_Anomaly_Score	0	1	0.505	0.270	0.520	0.280	0.730
Intrusion_Attempt	0	9	3.276	2.845	2	1	5
	Energy Source: Grid						
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1096.0	514.3	1102.9	648.7	1545.6
Energy_Consumed	152.1	1799.6	978.6	465.9	960.3	589.5	1383.3
Carbon_Emissions	10.05	199.7	79.36	51.74	59.25	38.56	117.2
Energy_Scheduling	0.600	0.990	0.798	0.104	0.795	0.714	0.884
Network_Traffic_Volume	1011	19978	10566.8	5462.0	10345	6119.5	15208.8
System_Anomaly_Score	0	1	0.505	0.270	0.520	0.280	0.730
Intrusion_Attempt	0	9	3.279	2.845	2	1	5
	Energy Source: Solar						
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1095.8	514.4	1102.5	648.4	1544.7
Energy_Consumed	152.1	1799.6	978.9	466.0	962.1	589.7	1383.3
Carbon_Emissions	10.05	199.66	79.39	51.75	59.27	38.57	117.3
Energy_Scheduling	0.600	0.990	0.798	0.104	0.794	0.714	0.884
Network_Traffic_Volume	1011	19978	10567.2	5460.8	10345	6122.5	15208.3
System_Anomaly_Score	0	1	0.505	0.270	0.520	0.280	0.730
Intrusion_Attempt	0	9	3.281	2.845	2	1	5
	Energy Source: Wind						
	Min.	Max.	Avg	Std	Median	1st quartile	3rd quartile
Energy_Generated	201.8	1997.7	1096.3	514.3	1102.8	649.4	1546.1
Energy_Consumed	152.1	1799.6	978.9	465.9	962.1	589.8	1383.3
Carbon_Emissions	10.05	199.7	79.36	51.74	59.27	38.57	117.2
Energy_Scheduling	0.600	0.990	0.798	0.104	0.795	0.714	0.885
Network_Traffic_Volume	1011	19978	10573.9	5460.5	10346	6124	15209
System_Anomaly_Score	0	1	0.505	0.270	0.520	0.280	0.730
Intrusion_Attempt	0	9	3.283	2.846	2	1	5

Table 2 compares the distributions of categorical variables in the original and balanced datasets after applying the Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-ENC). SMOTE-ENC was employed to address class imbalance while preserving the relationships between categorical and continuous features, thereby improving the representation of minority classes for

subsequent model training. As observed, the distributions of the energy source-related categories, including Load Balancing Status, become more balanced after preprocessing. Similarly, the frequencies of the Request Type categories increase, providing a more representative distribution of network requests. The Access Control Status and Blockchain Verification variables also exhibit improved class

representation while maintaining their original categorical structure. Likewise, the target variable, Intrusion Status, achieves a more balanced distribution, reducing class bias and

enabling the machine learning models to learn more effectively from both normal and intrusion instances.

Table 2. Distribution of Categorical Variables in Original and Balanced Datasets Across Energy Sources

Variable		Sample Number							
		Original				Balanced			
		Battery	Grid	Solar	Wind	Battery	Grid	Solar	Wind
Load _Balancing _Status	0	128	109	127	127	271	199	252	258
	1	107	143	129	110	188	229	230	200
	2	144	123	123	130	170	141	148	142
Request _Type	0	88	81	65	80	135	130	123	136
	1	77	71	74	73	166	113	142	149
	2	88	69	84	60	151	115	146	111
	3	66	86	78	70	112	132	132	114
	4	60	68	78	84	65	79	87	90
Access _Control _Status	0	191	177	186	178	441	371	437	411
	1	188	198	193	189	188	198	193	189
Blockchain _Verification	0	179	182	197	183	388	325	386	348
	1	200	193	182	184	241	244	244	252
Intrusion Status	0	72	64	75	75	322	258	326	308
	1	307	311	304	292	307	311	304	292

Figure 1 illustrates VIF values for the predictor variables, with the height of each bar indicating the level of multicollinearity for that feature. As shown in the figure, the VIF values are clustered around unity, ranging from 1.00253 for Access Control Status to 1.01375 for Load Balancing

Status. A VIF value of 1 implies that all predictors are independent in a model; however, values above 5 are considered critical for the presence of significant multicollinearity. As a result, all VIF values are below 5, indicating no multicollinearity in the dataset.

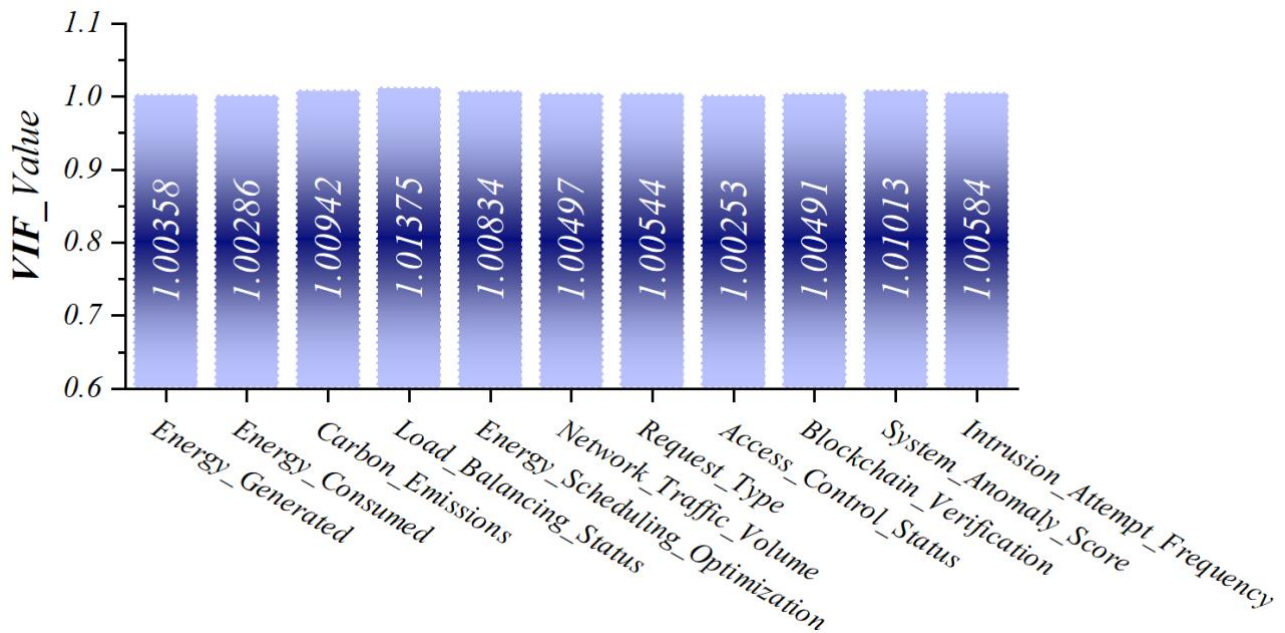


Fig. 1. VIF Analysis of Feature Independence and Multicollinearity

4. Result and Discussion

4.1. K- Fold Cross Validation

Table 3 presents the results of the five-fold cross-validation, in which the dataset was randomly partitioned into five equal folds. During each iteration, four folds (80% of the data) were used for training and one fold (20%) was used for testing, ensuring that every sample served as testing data exactly once. The models were evaluated using the F1-score across the five folds to assess their robustness and

generalizability. LDA demonstrates highly consistent performance, with F1-scores ranging from 0.7952 to 0.8168, yielding an average of 0.8040 and the lowest standard deviation (0.0094), indicating excellent stability but comparatively lower predictive capability. ETC achieves a higher average F1-score of 0.8128, with values ranging from 0.7854 to 0.8263 and moderate variability (Std. = 0.0159). HGBC delivers the best overall performance, achieving the highest average F1-score of 0.8235 despite slightly greater variability (Std. = 0.0202). Notably, HGBC attains the highest F1-score of 0.8465 in Fold K5, demonstrating its superior predictive capability among the evaluated models.

Table 3. K-Fold Cross-Validation Performance Comparison of Base Models using F1-Score

LDA		ETC		HGBC	
Fold	F1_Score	Fold	F1_Score	Fold	F1_Score
K1	0.8101	K1	0.7854	K1	0.8125
K2	0.8025	K2	0.8165	K2	0.8365
K3	0.7954	K3	0.8154	K3	0.7952
K4	0.7952	K4	0.8204	K4	0.8265
K5	0.8168	K5	0.8263	K5	0.8465
Avg.	0.8040	Avg.	0.8128	Avg.	0.8235
Std.	0.0094	Std.	0.0159	Std.	0.0202

4.2. Hyperparameter Results

The hyperparameters of the LDA, ETC, and HGBC models were optimized using the HROA and KOA algorithms. During the optimization process, each candidate

solution encoded a set of classifier hyperparameters sampled from the predefined search spaces listed in Table 4. Continuous hyperparameters were represented as real-valued

variables, whereas integer-valued hyperparameters ($n_estimators$, max_iter , $min_samples_split$, $min_samples_leaf$, and max_leaf_nodes) were rounded to the nearest valid integer before model training and evaluation. For each candidate solution, the corresponding classifier was trained and evaluated, and the F1-score was used as the fitness function to be maximized. The optimal hyperparameter values obtained for each model are summarized in Table 4. Specifically, the optimal tolerance for LDA was 1.78×10^{-4} ; the ETC model achieved the best performance with 132 estimators, a minimum samples split of 4, and a minimum samples leaf of 2; and the HGBC model was optimized with

200 maximum iterations, a learning rate of 0.20, and 34 maximum leaf nodes.

The optimization settings for HROA and KOA are presented in Table 5. To ensure a fair comparison, both algorithms were configured using the same population size (40 search agents), maximum number of iterations (200), stopping criterion (200 iterations), objective function (F1-score), solution encoding (continuous-valued vectors), and random seed (42). Using identical optimization settings ensured that any differences in classification performance resulted from the search strategies of HROA and KOA rather than differences in algorithm configuration.

Table 4. Optimized Hyperparameters for the models.

Model	Hyperparameter	Search Space	Optimized Value
LDA	Tolerance (tol)	$[1 \times 10^{-6}, 1 \times 10^{-2}]$	1.78×10^{-4}
ETC	Number of Estimators ($n_estimators$)	[50, 300]	132
	Minimum Samples Split ($min_samples_split$)	[2, 20]	4
	Minimum Samples Leaf ($min_samples_leaf$)	[1, 10]	2
HGBC	Maximum Iterations (max_iter)	[50, 300]	200
	Learning Rate ($learning_rate$)	[0.01, 0.30]	0.20
	Maximum Leaf Nodes (max_leaf_nodes)	[10, 100]	34

Table 5. Configuration of the HROA and KOA Optimization Algorithms

Parameter	HROA	KOA
Population Size	40 search agents	40 search agents
Maximum Iterations	200	200
Random Seed	42	42
Stopping Criterion	200	200
Objective Function	F1-Score	F1-Score
Solution Encoding	Continuous-valued vector	Continuous-valued vector

4.3. Performance Evaluation of Machine Learning Models

Table 6 compares the predictive performance of the baseline classifiers and their HROA- and KOA-optimized counterparts on the training and test datasets using Accuracy, Precision, Recall, F1-score, MCC, Markedness, AUC, Brier Binary Score, and runtime. Among the baseline models, HGBC achieves the highest training and test accuracies of 0.854 and 0.835, respectively, followed by ETC and LDA. HGBC also records the lowest Brier Binary Score (0.150) and the highest AUC (0.853) among the baseline models, indicating superior discrimination ability and probability calibration. The baseline models exhibit relatively short execution times, requiring approximately 12–13 s.

Hyperparameter optimization substantially improves the performance of all classifiers. Among the hybrid models, HGBC(H1) demonstrates the best overall performance, achieving the highest test Accuracy (0.942), Precision (0.952),

Recall (0.942), F1-score (0.947), MCC (0.900), Markedness (0.900), and AUC (0.952), while simultaneously obtaining the lowest Brier Binary Score (0.050), indicating excellent predictive reliability and calibration. ETC(H1) ranks second, with a test Accuracy of 0.926, F1-score of 0.924, and Brier Binary Score of 0.070, whereas LDA(H1) also exhibits considerable improvement over the baseline model, achieving a test Accuracy of 0.899 and reducing the Brier Binary Score to 0.090. Although the optimized models require longer execution times (113–125 s) because of the optimization process, the substantial improvements in predictive accuracy, calibration, and classification performance demonstrate the effectiveness of integrating HROA and KOA with the baseline classifiers. Overall, HGBC(H1) provides the best trade-off between predictive performance and model reliability, making it the most effective model for intrusion prediction in renewable energy-powered data center environments.

Table 6. Comparative Performance Metrics of Base and Hybrid Models Across Training and Test

		values			values			values	
	Model: LDA	Train	Test	Model: LDA(H1)	Train	Test	Model: LDA(H1)	Train	Test
Accuracy		0.825	0.800		0.912	0.899		0.903	0.887
Precision		0.884	0.815		0.939	0.902		0.932	0.891
Recall		0.820	0.800		0.900	0.900		0.889	0.887
F1_core		0.851	0.807		0.919	0.901		0.909	0.889
MCC		0.636	0.640		0.815	0.820		0.795	0.800
Markedness		0.625	0.641		0.810	0.821		0.790	0.801
AUC		0.820			0.910			0.900	
Run time		12s			114s			116s	
Brier Binary		0.180			0.090			0.100	
		values			values			values	
	Model: ETC	Train	Train	Model: ETC(H1)	Train	Test	Model: ETC(H2)	Train	Test
Accuracy		0.834	0.813		0.931	0.926		0.895	0.870
Precision		0.891	0.825		0.955	0.927		0.927	0.880
Recall		0.838	0.832		0.933	0.921		0.904	0.872
F1_core		0.863	0.828		0.944	0.924		0.915	0.876
MCC		0.655	0.660		0.854	0.860		0.778	0.780
Markedness		0.645	0.660		0.848	0.860		0.780	0.780
AUC		0.830			0.930			0.890	
Run time		12s			113s			118s	
Brier Binary		0.170			0.070			0.110	
		values			values			values	
	Model: HGBC	Train	Test	Model: HGBC(H1)	Train	Test	Model: HGBC(H2)	Train	Test
Accuracy		0.854	0.835		0.952	0.942		0.921	0.918
Precision		0.780	0.855		0.923	0.952		0.880	0.924
Recall		0.849	0.835		0.952	0.942		0.913	0.918
F1_core		0.813	0.845		0.937	0.947		0.896	0.921
MCC		0.695	0.700		0.899	0.900		0.832	0.840
Markedness		0.685	0.700		0.893	0.900		0.827	0.841
AUC		0.853			0.952			0.919	
Run time		13s			125s			122s	
Brier Binary		0.150			0.050			0.080	

4.4. Analysis of the Wilcoxon Test

Table 7 presents the results of the pairwise Wilcoxon signed-rank tests comparing the base and hybrid models, including the test statistics and corresponding *p*-values. For

each pairwise comparison, the paired observations consisted of the prediction results obtained for the same samples in the complete dataset, ensuring that both models were evaluated on identical observations. The statistical analysis was performed using the prediction results generated from a single

optimization run with a fixed random seed of 42. No multiple-comparison correction was applied, as the Wilcoxon tests were conducted as exploratory pairwise analyses to complement the comparative performance evaluation.

The results indicate that most pairwise comparisons produced p -values greater than the significance level of 0.05, suggesting that the observed performance differences between the corresponding models are not statistically significant. This pattern is evident for the LDA model and its hybrid variants, as well as for most comparisons involving the ETC and HGBC

models. Nevertheless, several comparisons yielded statistically significant differences, including ETC(H1) versus ETC(H2) ($p = 0.0093$) and HGBC(H1) versus HGBC(H2) ($p = 0.0140$). Overall, the statistical analysis indicates that, although hybrid optimization occasionally improves classification performance, the majority of the observed improvements are modest and not statistically significant. Among the evaluated classifiers, the HGBC-based variants demonstrate the most consistent performance improvements relative to their corresponding baseline model.

Table 7. Wilcoxon Signed-Rank Test Results Comparing Base and Hybrid Models

LDA:1, LDA(H1):2, LDA(H2):3 ETC:4, ETC(H1):5, ETC(H2):6 HGBC:7, HGBC(H1):8, HGBC(H2):9											
1+	Statistic	P-Value	2+	Statistic	P-Value	3+	Statistic	P-Value	4+	Statistic	P-Value
2	11716.5	0.7865									
3	9067.5	0.5657	3	125	0.4142						
4	137.5	0.6831	4	9165	0.6666	4	6840	0.4431			
5	17018	0.4263	5	400	0.0152	5	962	0.0140	5	14152	0.4804
6	6754.5	0.3574	6	490	0.2482	6	125	0.4142	6	4851	0.2466
7	1258	0.5584	7	4964	0.4548	7	3294	0.2373	7	575	0.6682
8	23775	0.3681	8	1930.5	0.0434	8	3013.5	0.0298	8	20363.5	0.4126
9	14274	0.5637	9	78	0.0093	9	400	0.0152	9	11660	0.6362
5+	Statistic	P-Value	6+	Statistic	P-Value	7+	Statistic	P-Value	8+	Statistic	P-Value
6	1764	0.0111									
7	9067.5	0.5657	7	1960	0.0843						
8	575	0.6682	8	4336.5	0.0205	8	14152	0.4804			
9	125	0.4142	9	962	0.0140	9	7096.5	0.7590	9	1221	0.4126

4.5. Performance Comparison of Base and Hybrid Models

Confusion matrices for the baseline and optimized LDA, ETC, and HGBC models on the test dataset are provided in Figure 2. Since the test set includes only negative examples, the presented matrices show the ability of each model to correctly classify negative samples without any false alarms. All baseline models have rather high values of the false positive error, with 97, 91, and 80 examples misclassified by LDA, ETC, and HGBC, correspondingly. However, the use of

the optimization algorithms decreased significantly the number of erroneous classifications for all families of models. Out of all optimized models, HROA algorithm managed to decrease the number of false positives for LDA, ETC, and HGBC to 49, 36, and 28, correspondingly, while increasing at the same time the number of correctly classified negative examples up to 437, 450, and 458. KOA algorithm helped to improve the baseline models as well, although its performance was worse than that of HROA in every case. In summary, HGBC+HROA proved to be the most effective combination with the best generalization performance; it showed the

minimum number of false positives (28) and the maximum number of correctly classified negative examples (458 out of 486).

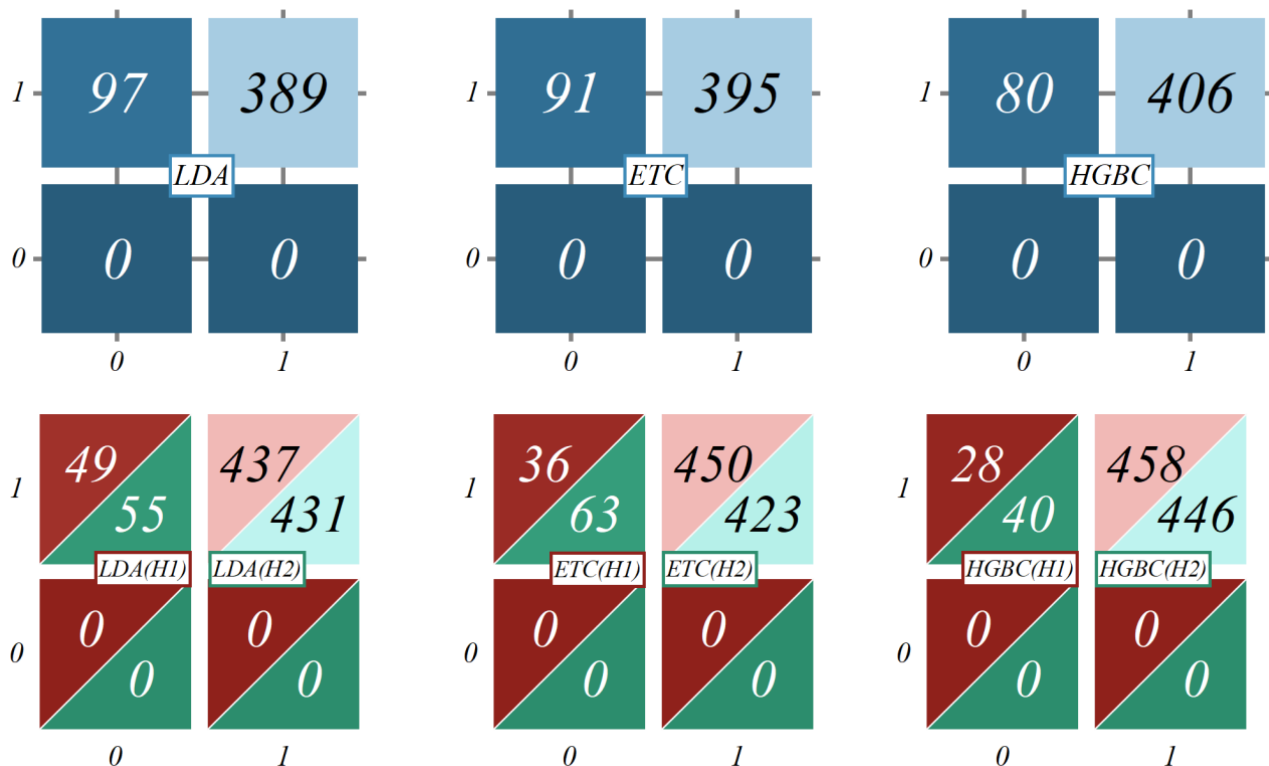


Fig. 2. Confusion Matrix Analysis of Base and Hybrid Models

4.6. Sensitivity Analysis

In order to determine the significance of the input variables and their interrelations, Copula-based sensitivity analysis was conducted, as shown in Figure 3 and Table 8 & 9. The values of maximum mutual information (MI) were calculated for Intrusion Attempt Frequency (0.416), Energy Scheduling Optimization (0.327), Access Control Status (0.312) and System Anomaly Score (0.278). This shows that cybersecurity and operational control variables provide the most significant information for predicting system anomaly. On the other hand, traditional energy-related variables like Energy Generated (0.018) and Energy Consumed (0.006) have lower predictive significance as they reflect normal behavior of the system instead of anomalies caused by cyber attacks, which are sought to be identified using the proposed system detection model.

This pattern of results is supported by the correlation analysis presented in Table 9 below. The System Anomaly Score exhibits moderate positive correlations with Intrusion Attempt Frequency ($r = 0.52$), whereas Intrusion Status exhibits a strong correlation with Access Control Status ($r =$

0.67) and moderate correlation with Intrusion Attempt Frequency ($r = 0.52$). Therefore, these dependencies imply that cyberattacks impact several components of the system at once and are therefore dependable indicators of potential security breaches that could disrupt the functioning of renewable-energy systems.

Despite the fact that the energy generation and consumption variables play lesser roles in the classification process, their importance for the functioning of renewable-energy systems remains unchanged. Instead, the analysis reveals that the proposed framework recognizes cyber threats by detecting anomalies in security and control mechanisms of the system before they propagate to the energy infrastructure level. Such early recognition of threats allows implementing mitigation measures which protect the integrity of energy generation processes, sustain energy scheduling and load balancing functions, prevent disruptions of services, and improve the overall reliability and resilience of renewable-energy systems in case of a cyberattack. Thus, the sensitivity analysis proves the importance of cybersecurity monitoring as the enabling factor in ensuring the protection of renewable-energy systems.

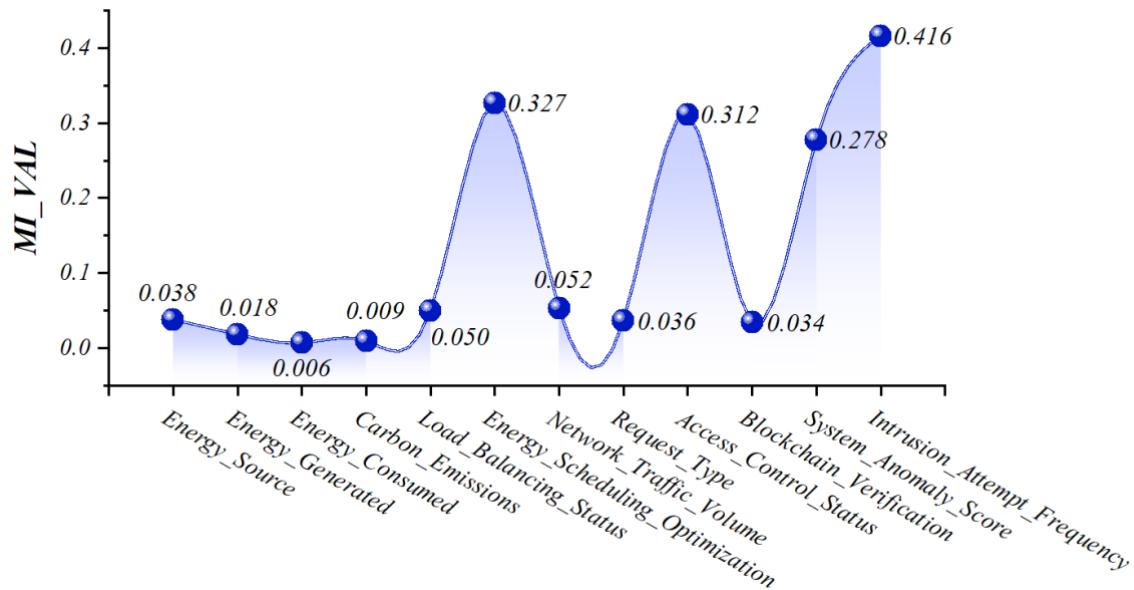


Fig. 3. Sensitivity Analysis of System Variables via Copula-Based Mutual Information Distribution

Table 8. List of System Variables and Features in Copula

1	Energy_Source
2	Energy_Generated
3	Energy_Consumed
4	Carbon_Emissions
5	Load_Balancing
6	Energy_Scheduling_Optimization
7	Network_Traffic_Volume
8	Request_Type
9	Access_Control_Status
10	Blockchain_Verification
11	System_Anomaly_Score
12	Intrusion_Attempt_Frequency
13	Intrusion Status

Table 9. Pairwise Correlation Coefficients Among System Variables in Copula

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	-0.01	0.04	-0.50	-0.02	0.00	0.00	0.05	-0.01	0.01	-0.01	0.01	0.00
2	-0.01	1	0.03	0.03	0.01	0.01	-0.04	0.02	-0.03	-0.03	-0.01	-0.02	-0.03
3	0.04	0.03	1	-0.02	-0.01	0.06	0.02	-0.04	0.00	-0.01	-0.03	0.00	0.00
4	-0.50	0.03	-0.02	1	0.04	0.02	-0.06	-0.05	0.00	0.01	0.04	0.01	-0.01
5	-0.02	0.01	-0.01	0.04	1	-0.10	-0.03	0.09	0.24	0.09	0.05	0.17	0.20
6	0.00	0.01	0.06	0.02	-0.10	1	-0.03	0.00	-0.01	0.00	-0.02	-0.01	-0.04
7	0.00	-0.04	0.02	-0.06	-0.03	-0.03	1	0.00	0.00	-0.01	-0.06	-0.05	-0.02

8	0.05	0.02	-0.04	-0.05	0.09	0.00	0.00	1	0.10	0.08	0.01	0.12	0.11
9	-0.01	-0.03	0.00	0.00	0.24	-0.01	0.00	0.10	1	0.23	0.01	0.35	0.67
10	0.01	-0.03	-0.01	0.01	0.09	0.00	-0.01	0.08	0.23	1	-0.02	0.20	0.24
11	-0.01	-0.01	-0.03	0.04	0.05	-0.02	-0.06	0.01	0.01	-0.02	1	0.04	-0.01
12	0.01	-0.02	0.00	0.01	0.17	-0.01	-0.05	0.12	0.35	0.20	0.04	1	0.52
13	0.00	-0.03	0.00	-0.01	0.20	-0.04	-0.02	0.11	0.67	0.24	-0.01	0.52	1

5. Conclusion

The present digital infrastructure relies heavily on energy sources, especially in carbon-neutral data centers that integrate the grid, solar, wind, and battery power to support sustainability and business continuity. In addition to reducing carbon emissions, the production of renewable energy improves cost-effectiveness and sustainability. However, infrastructure reliant on renewable energy sources is becoming increasingly vulnerable to cybersecurity threats as digitization and networking grow. Consequently, predicting the intrusion status of infrastructure is essential to maintaining its stability. Predicting intrusion status effectively safeguards the security of energy management systems, so Machine Learning (ML) has been crucial for uncovering hidden patterns in complex data. The three main classification models used in this study were Histogram-Based Gradient Boosting Classification (HGBC), Extra Trees Classifier (ETC), and Linear Discriminant Analysis (LDA), which were optimized by the Heavy Rain Optimization Algorithm (HROA) and the Kangaroo Optimization Algorithm (KOA) to improve their performance. Before developing the model, multicollinearity was assessed using VIFs, which showed a decreasing pattern, with an initial value of 5. The absence of multicollinearity is indicated by a VIF value of 1. Additionally, to balance the dataset, SMOTE-ENC oversampling was used, and to validate the model, K-fold cross-validation was used. The findings revealed that HGBC outperformed LDA and ETC in both cross-validation and other evaluations. In the 5-fold validation, HGBC achieved the highest F1 score (mean 0.8235), but LDA demonstrated the highest stability. Similarly, in the training set, HGBC achieved the highest Accuracy (0.854) and the lowest Brier score (0.150). On the test set, hybrid models performed better than baseline models; specifically, HGBC(H1) showed the best test accuracy (0.942)

and the minimum Brier Binary score (0.050). Although the Wilcoxon signed-rank test showed that most performance differences were not significant ($p > 0.05$), hybrid models still showed performance advantages. In a Copula-based sensitivity analysis, it was found that security-related features, Intrusion Attempt Frequency and Access Control Status, were more important than others. Subsequently, the recommended hybrid model is HGBC(H1), which was optimized using HROA due to its high predictive ability, excellent generalization performance, and the lowest Brier Score of 0.050. From the comparative results, it is evident that the HROA-based optimizations performed better and more stably across all models than KOA, thereby demonstrating the effectiveness of the HROA optimization technique for predicting intrusions in data centers with integrated renewable energy systems.

Despite the promising results obtained in this study, several limitations should be acknowledged. The proposed framework was evaluated using a single benchmark dataset of limited size, which may not fully capture the diversity and complexity of real-world renewable-energy infrastructures and evolving cyberattack scenarios. Another limitation of the present study lies in the preprocessing strategy adopted for class balancing. SMOTE-ENC was applied to generate a balanced dataset, enabling a comparative statistical analysis between the original and balanced datasets, and facilitating the subsequent model development. However, the oversampling procedure was not integrated within each training fold of the cross-validation framework. Although this approach allowed an effective assessment of the impact of data balancing, future studies will implement fold-wise SMOTE-ENC within a leakage-free cross-validation pipeline to further improve methodological rigor and provide a more conservative estimate of model performance.

Abbreviations

Definition	Symbol
Machine Learning	ML
Linear Discriminant Analysis	LDA
Extra Trees Classifier	ETC
Histogram-Based Gradient Boosting Classification	HGBC
Heavy Rain Optimization Algorithm	HROA
Kangaroo Optimization Algorithm	KOA

LDA+HROA	LDA(H1)
LDA+KOA	LDA(H2)
ETC+HROA	ETC(H1)
ETC+KOA	ETC(H2)
HGBC+HROA	HGBC(H1)
HGBC+KOA	HGBC(H2)
Variance Inflation Factor	VIF
Synthetic Minority Oversampling Technique Encoded Nominal and Continuous	SMOTE-ENC
True Positive	TP
True Negative	TN
False Positive	FP
False Negative	FN
Total number of samples or predictions	N
True Positive Rate	TPR
False Positive Rate	FPR
Predicted probability of the positive class	P_i
Observed outcome	O_i
Proportion of correctly classified instances	Accuracy
Ratio of true positives to predicted positives	Precision
Ratio of true positives to actual positives	Recall
Harmonic mean of Precision and Recall	F1-score
Measure of class discriminative power	Markedness
Error rate computed separately for each class	Class-wise Error
Mean-squared error of probabilistic predictions	Brier Score
Matthews Correlation Coefficient	MCC
Area Under the Receiver Operating Characteristic Curve	AUC

Competing of interests

The authors declare no competing of interests.

Authorship Contribution Statement

Min ZHANG: Writing-Original draft preparation, Conceptualization, Project administration.

Data Availability

The original dataset is publicly available in <https://www.kaggle.com/datasets/zara2099/renewable-cyber-defense-dataset>.

Declarations

Not applicable

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Funding

Not applicable

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

References

- [1] Ł. Mazur, S. Cieřlik, and S. Czapp, "Trends in locally balanced energy systems without the use of fossil fuels: A review," *Energies (Basel)*, vol. 16, no. 12, p. 4551, 2023.
- [2] C.-W. Ten and Y. Hou, *Modern power system analysis*. CRC Press, 2024.
- [3] E. Hrayshat, "Techno-Economic Analysis of an Autonomous Hybrid PV-Biomass-Battery Power System," *International Journal of Renewable Energy Research*, 2025, doi: 10.20508/ijrer.v15i4.15031.g9126.
- [4] A. Kalair, N. Abas, M. S. Saleem, A. R. Kalair, and N. Khan, "Role of energy storage systems in energy transition from fossil fuels to renewables," *Energy Storage*, vol. 3, no. 1, p. e135, 2021.
- [5] T. H. Mehedi, E. Gemechu, and A. Kumar, "Life cycle greenhouse gas emissions and energy footprints of utility-scale solar energy systems," *Appl. Energy*, vol. 314, p. 118918, 2022.
- [6] J.-N. Kang, Y.-M. Wei, L.-C. Liu, R. Han, B.-Y. Yu, and J.-W. Wang, "Energy systems for climate change mitigation: A systematic review," *Appl. Energy*, vol. 263, p. 114602, 2020.
- [7] P. A. Østergaard, N. Duic, Y. Noorollahi, and S. Kalogirou, "Renewable energy for sustainable development," *Renew. Energy*, vol. 199, pp. 1145–1152, 2022.
- [8] M. A. Basit, S. Dilshad, R. Badar, and S. M. Sami ur Rehman, "Limitations, challenges, and solution approaches in grid-connected renewable energy systems," *Int. J. Energy Res.*, vol. 44, no. 6, pp. 4132–4162, 2020.
- [9] A. Harrouz, D. Belatrache, K. Boulal, I. Colak, and K. Kayisli, "Social Acceptance of Renewable Energy dedicated to Electric Production," in *2020 9th International Conference on Renewable Energy Research and Application (ICRERA)*, 2020, pp. 283–288. doi: 10.1109/ICRERA49962.2020.9242904.
- [10] M. V Dambhare, B. Butey, and S. V Moharil, "Solar photovoltaic technology: A review of different types of solar cells and its future trends," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 12053.
- [11] P. Gajewski and K. Pieńkowski, "Control of the hybrid renewable energy system with wind turbine, photovoltaic panels and battery energy storage," *Energies (Basel)*, vol. 14, no. 6, p. 1595, 2021.
- [12] S. I. Kirici and H. Yurtsever, "Renewable Energy, Economic Growth, and an Examination Related to Wind Energy: The Case of Selected OECD Countries," *International Journal of Renewable Energy Research*, 2025, doi: 10.20508/ijrer.v15i4.15077.g9144.
- [13] N. H. S. A. Z. Yasser Elsayed, "Assessing Wind Energy Conversion Systems Based on Newly Developed Wind Turbine Emulator," *International Journal of Smart grid*, vol. 4, no. 4, Dec. 2020, doi: 10.20508/ijsmartgrid.v4i4.133.g101.
- [14] C. Koç, "A study on solutions and problems of hydroelectric power plants in the operation," *Sustain. Water Resour. Manag.*, vol. 8, no. 3, p. 90, 2022.
- [15] M. R. A. Bhuiyan, "Overcome the future environmental challenges through sustainable and renewable energy resources," *Micro Nano Lett.*, vol. 17, no. 14, pp. 402–416, 2022.
- [16] Guler Ilkay, "The Transition of CO2 Emission and Employment Multipliers from Brown to Green Under a Renewable Energy-Based Economic Growth Scenario in Türkiye:2053 Vision," *International Journal of Smart grid*, vol. 9, no. v9i2, 2025, doi: 10.20508/ijsmartgrid.v9i2.373.g379.
- [17] X. Xianyong and Z. Zixuan, "New Power Systems Dominated by Renewable Energy Towards the Goal of Emission Peak & Carbon Neutrality: Contribution, Key Techniques, and Challenges," *Advanced Engineering Science/Gongcheng Kexue Yu Jishu*, vol. 54, no. 1, 2022.
- [18] S. Saeed and T. Siraj, "Global renewable energy infrastructure:: pathways to carbon neutrality and sustainability," *Solar energy and sustainable development journal*, vol. 13, no. 2, pp. 183–203, 2024.
- [19] D. E. Ekechukwu and P. Simpa, "The future of Cybersecurity in renewable energy systems: A review, identifying challenges and proposing strategic solutions," *Computer Science & IT Research Journal*, vol. 5, no. 6, pp. 1265–1299, 2024.
- [20] A. Rekeraho, D. T. Cotfas, P. A. Cotfas, T. C. Bălan, E. Tuyishime, and R. Acheampong, "Cybersecurity challenges in IoT-based smart renewable energy," *Int. J. Inf. Secur.*, vol. 23, no. 1, pp. 101–117, 2024.
- [21] J. P. Yaacoub, H. Noura, J. Azar, O. Salman, and K. Chahine, "Cybersecurity in smart renewable energy systems," in *2024 International Wireless Communications and Mobile Computing (IWCMC)*, IEEE, 2024, pp. 1534–1540.
- [22] H.-C. Lin, P. Wang, K.-M. Chao, W.-H. Lin, and Z.-Y. Yang, "Ensemble learning for threat classification in network intrusion detection on a security monitoring system for renewable energy," *Applied Sciences*, vol. 11, no. 23, p. 11283, 2021.
- [23] F. Almutairy, L. Scekkic, R. Elmoudi, and S. Wshah, "Accurate detection of false data injection attacks in renewable power systems using deep learning," *Ieee Access*, vol. 9, pp. 135774–135789, 2021.
- [24] U. Islam, H. Ullah, N. Khan, K. Saleem, and I. Ahmad, "AI-enhanced intrusion detection in smart renewable energy grids: A novel industry 4.0 cyber threat

- management approach,” *International Journal of Critical Infrastructure Protection*, vol. 50, p. 100769, 2025.
- [25] M. M. Forootan, I. Larki, R. Zahedi, and A. Ahmadi, “Machine learning and deep learning in energy systems: A review,” *Sustainability*, vol. 14, no. 8, p. 4832, 2022.
- [26] O. A. Alimi, K. Ouahada, A. M. Abu-Mahfouz, and K. O. A. Alimi, “Empirical comparison of machine learning algorithms for mitigating power systems intrusion attacks,” in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, IEEE, 2020, pp. 1–5.
- [27] K. Gokulraj and C. B. Venkatramanan, “Advanced machine learning-driven security and anomaly identification in inverter-based cyber-physical microgrids,” *Electric Power Components and Systems*, pp. 1–18, 2024.
- [28] S. Ghyasuddin Hashmi *et al.*, “Machine learning-based renewable energy systems fault mitigation and economic assessment,” *Electric Power Components and Systems*, pp. 1–24, 2024.
- [29] K. Aygul, M. Mohammadpourfard, M. Kesici, F. Kucuktezcan, and I. Genc, “Benchmark of machine learning algorithms on transient stability prediction in renewable rich power grids under cyber-attacks,” *Internet of Things*, vol. 25, p. 101012, 2024.
- [30] S. Zhao, B. Zhang, J. Yang, J. Zhou, and Y. Xu, “Linear discriminant analysis,” *Nature Reviews Methods Primers*, vol. 4, no. 1, p. 70, 2024.
- [31] A. Sharaff and H. Gupta, “Extra-tree classifier with metaheuristics approach for email classification,” in *Advances in computer communication and computational sciences: proceedings of IC4S 2018*, Springer, 2019, pp. 189–197.
- [32] H. Nhat-Duc and T. Van-Duc, “Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification,” *Autom. Constr.*, vol. 148, p. 104767, 2023.
- [33] J. Zhang, “Kangaroo Optimization Algorithm.”
- [34] S. Y. Park, J. E. Park, H. Kim, and S. H. Park, “Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches),” vol. 22, no. 10, pp. 1697–1707, 2021.
- [35] I. K. Nti, O. Nyarko-Boateng, and J. Aning, “Performance of machine learning algorithms with different K values in K-fold cross-validation,” *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, 2021.
- [36] N. I. S. Indah and F. Ahmad, “Exploring the Wilcoxon Test in Science Education: A Literature Review of Empirical Research,” *Indonesian Journal of Educational Science (IJES)*, vol. 7, no. 2, pp. 158–169, 2025.
- [37] G. Divine, H. J. Norton, R. Hunt, and J. Dienemann, “A review of analysis and sample size calculation considerations for Wilcoxon tests,” *Anesth. Analg.*, vol. 117, no. 3, pp. 699–710, 2013.
- [38] M. Mukherjee and M. Khushi, “SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features,” *Applied System Innovation*, vol. 4, no. 1, p. 18, 2021.
- [39] P. Wei, Z. Lu, and J. Song, “Moment-independent sensitivity analysis using copula,” *Risk Analysis*, vol. 34, no. 2, pp. 210–222, 2014.
- [40] C.-T. Yeh, G.-Y. Liao, and T. Emura, “Sensitivity analysis for survival prognostic prediction with gene selection: a copula method for dependent censoring,” *Biomedicines*, vol. 11, no. 3, p. 797, 2023.
- [41] “Renewable Cyber Defense Dataset.” [Online]. Available: <https://www.kaggle.com/datasets/zara2099/renewable-cyber-defense-dataset>