Understanding Energy Saving Potential In Built And Industrial Systems For Enabling Cost Reduction And Environmental Sustainability

Jing Shi

Department of Engineering Management, Henan Technical College of Construction, Zhengzhou 450064, Henan, China

Corresponding author. E-mail: fwzyqsl@126.com

Received: Sep. 29, 2025; Accepted: July. 04, 2026

Energy conservation emerged as a global priority due to rising energy demand, environmental concerns, and the share of industrial and urban sectors in consumption. Energy-saving potential requires exact prediction because this information enables the creation of efficient retrofit methods and sustainable energy policy frameworks. This study develops and evaluates predictive models for estimating energy-saving potential using a dataset of 2,191 aggregated daily samples derived from 52,584 raw measurements. The dataset includes 29 input variables, with feature selection performed through the Variance Inflation Factor (VIF) analysis to address multicollinearity. Three Machine Learning (ML) models, Extra Trees Regression (ETR), Quantile Regression (QR), and Stochastic Gradient Boosting (SGB), are implemented and optimized using two metaheuristic algorithms, the Kepler Optimization Algorithm (KOA) and the Gazelle Optimization Algorithm (GOA). Model performance is assessed via 5-fold cross-validation and evaluated using R^2, RMSE, COV, PI, MAE, and MARD metrics, along with runtime analysis. Statistical robustness is ensured through the Wilcoxon test, while SHAP analysis was performed on the best-performing model to identify the most influential features driving energy-saving potential. The best-performing model was SGKA (KOA-optimized SGB), which achieved the highest accuracy (R^2 = 0.975) and the lowest RMSE (0.162), along with the smallest MAE and MARD values in the test phase. Optimized ensemble learning methods produce both precise and dependable, and easily understood prediction results, according to the research findings. The research field will progress through the development of hybrid physics–data systems and deep learning methods, and real-time prediction systems. These will include occupant behavior analysis.

Keywords: Energy-saving potential; Machine learning; Extra trees regression; Quantile regression; Wilcoxon test

http://dx.doi.org/10.6180/jase.202611_34.001

1. Introduction

Over the past few years, energy conservation has attracted influential attention from researchers, resulting from growing concerns about the quick decline in energy sources, ecological problems [1], rising energy prices [2], and the increasing global energy demand [3]. As a result, reducing energy consumption is regarded as a crucial step toward ensuring sustainable energy use in urban environments [4]. Urban areas consume about 70% of global energy resources, according to the source [5], as their current growth patterns for cities and factories have boosted their power needs. The worldwide manufacturing industry has established energy efficiency improvement as its most vital goal for achievement. As climate change intensifies, the requirement to decrease and limit emissions has become increasingly pressing [6, 7]. The industrial sector alone accounts for approximately 30% of the world's total energy allocation [8]. Energy audits play a vital role as an initial strategy to cut down industrial energy use by analyzing

consumption patterns and recommending more productive practices [9, 10]. Home-based electricity usage increased because people now use refrigerators and washing machines, and mobile devices more frequently. To address this, optimizing energy use is necessary, and this study applies ML techniques for energy forecasting to help mitigate excessive power consumption.

ML employs data-driven models that focus on the correlations within the initial data and elements, rather than relying on equations and calculations [11]. Yang et al. [12] reviewed large-scale building retrofit energy-saving prediction models, highlighting data-driven, physics-based, and hybrid approaches. They found that most models validate only at aggregate levels, ignore rebound/prebound effects, and exclude occupant willingness, limiting accuracy. Shen et al. [13] developed a bottom-up model for regional residential buildings in New York using RECS data and SimBldPy simulation. Optimization showed 48–62% primary energy savings potential by 2040 across different building types. Ko et al. [14] proposed a prediction-based preheating approach for coil control for energy recovery ventilators (ERVs). The method reduced energy use by 7–72% and decreased coil size by up to 21%, while preventing frost formation more efficiently than conventional methods. Xu et al. [15] introduced a prediction-based energy-saving scheme (PES) for object-tracking sensor networks. Simulations showed that PES significantly reduces energy use under varying system workloads and object behaviors. Chung and Burnett [16] modeled lighting energy savings with occupancy sensors, accounting for user-defined probabilities and random occupancy patterns. Demonstrated how different delay settings affect savings in retrofit programs. Kleiminger et al. [17] built a simulation framework based on ISO 13790 for evaluating occupancy-prediction-based heating controls. Showed predictive thermostat control can improve energy savings by tailoring heating to weather and occupancy patterns. Cho et al. [18] proposed a hydraulic balancing methodology that ensures thermal comfort while saving 2–14% energy in space heating. A Geneva case study confirmed that balancing improves both comfort and efficiency. Wittchen et al. [19] analyzed Denmark's building stock via energy certification data. A 30% national heating energy savings potential remains despite long-standing policies, emphasizing retrofits as key to achieving 2050 carbon neutrality. Yusop [20] designed a predictive cut-off control for pneumatic actuators, saving up to 80.2% of wasted air energy compared to conventional actuation, albeit with slightly longer actuation times. Widayati [21] showed that accounting for interdependency effects among energy-saving measures increases predicted savings in office retrofits (e.g., lighting and VSD retrofits). Improved estimates enhance investment confidence in energy projects. Zhang [22] applied ML (Random Forest, LightGBM) to analyze energy utilization in buildings. Results showed improved prediction accuracy and identified key parameters, with potential for future data integration.

The main purpose of this research is to upgrade and measure predictive models for estimating energy-saving potential using a dataset of 2,191 aggregated daily samples (from 52,584 raw measurements averaged over 24-hour intervals). The dataset contains 29 input variables (excluding time), together with one output variable, which shows the potential for energy saving. The VIF method performs feature selection by using a 10-point threshold to eliminate variables. These show multicollinearity. The research uses three ML models, which include ETR, QR, and SGB, together with two metaheuristic optimization methods called KOA and GOA. The hybrid models were created by combining each baseline prediction model with an optimization algorithm for hyperparameter tuning, where the short forms indicate both the base model and the optimizer used. Specifically, QRKA and QRGA represent QR optimized with KOA and GOA, respectively; ETKA and ETGA denote ETR optimized with KOA and GOA; and SGKA and SGGA correspond to SGB optimized with KOA and GOA, respectively. Model evaluation is conducted using 5-fold cross-validation, and performance is assessed through R^2, RMSE, COV, PI, MAE, and MARD metrics, along with the computation of run time. Statistical robustness is verified using the Wilcoxon test, while SHAP analysis was performed on the best-performing model to identify the most influential features. The aim is to identify a high-performing modeling strategy for predicting energy-saving potential while addressing issues of accuracy, stability, and interpretability.

2. Materials and methods

2.1. Data Collection

The dataset used in this research was gathered from the Kaggle site [23] and includes detailed energy consumption records from a diverse set of over 100 buildings in Southern California, spanning the period from January 2018 to January 2024. Multiple data sources, including smart meters, IoT sensors, building management systems, and regional utility companies, were utilized to collect data, ensuring high-resolution, real-world, and comprehensive coverage of electricity usage under various operational and environmental conditions. The dataset comprises hourly samples for residential, commercial, and industrial build-

ings, including both energy consumption measurements and contextual variables that influence demand patterns and efficiency. Besides overall electricity usage, the dataset includes submetered data for lighting and HVAC systems, as well as building operational parameters like occupancy rates and peak demand indicators. The study included environmental factors, which consisted of temperature and humidity, wind speed, solar radiation, and rainfall, to assess how weather conditions affect power consumption. The analysis included energy-related data, economic information, and sustainability metrics, which contained energy prices and carbon emission rates and categories for carbon emission reduction. To facilitate easy identification of records by seasonal cycles, holiday effects, and weekday versus weekend differences, a timestamp was assigned to each record, thereby clarifying the period under study. Hence, the final dataset presents a four-dimensional representation of energy consumption, encompassing physical, behavioral, economic, and environmental factors. The variable that corresponds to the target in this research is Energy Savings Potential, which is computed from the raw hourly measurements (52,584 records) by first aggregating them into daily totals (2,191 samples) and then calculating the potential savings as the difference between the actual daily energy consumption and the estimated minimum consumption under optimized operating conditions (derived from sub-metered lighting and HVAC data together with contextual and environmental variables using standard efficiency benchmarks). Its unit is kWh/day. This metric is a valid proxy for "savings potential" because it directly quantifies the actionable gap between current usage and theoretically achievable efficient usage, rather than being simply a transformed or scaled version of the total consumption metric. This characteristic, along with the current consumption patterns and the related contextual variables, is the predictive output of the ML models created in this research.

2.1.1. *Variance Inflation Factor (VIF) Feature Selection*

VIF feature selection is a statistical method that identifies and removes multicollinearity, a condition where predictor variables in a dataset are closely interconnected, resulting in distorted model estimates, increased variance of regression coefficients, and reduced interpretability and reliability of results, among other issues. High intercorrelation among predictors occurs when independent variables are highly correlated with each other, which can lead to model estimates being distorted, the variance of regression coefficients being significantly increased, and results losing their interpretability and reliability. VIF measures the amount of the variance that a regression coefficient increases because of collinearity. The VIF value that is highest mathematically means the strongest multicollinearity. To determine if a problem exists, typically, if the VIF of a variable is larger than 10, the variable is suggested for removal. In feature selection, VIF helps retain only the predictors that contribute independently to the model, improving both stability and interpretability [24].

Table 1 illustrates the results of the VIF feature selection process, showing both the initial and final VIF scores for each predictor. Initially, many variables showed excessively high VIF scores, showing considerable multicollinearity that could compromise model stability and interpretation. As part of the selection process, such highly correlated variables were removed, leaving only predictors with acceptable VIF values, typically below the common threshold of 10. The final retained variables include Temperature (1.631), Occupancy Rate (1.428), Building Age (5.737), Local Energy Production (6.557), and Smart Plug Usage (8.091), all of which demonstrate low collinearity and thus contribute independent predictive power to the model. This refinement ensures that the final dataset is both statistically robust and interpretable, reducing redundancy while maintaining the most relevant predictors for reliable analysis.

2.2. Methodology

2.2.1. *Extra Trees Regression (ETR)*

The ETR is an ensemble learning technique built on decision trees, designed for regression tasks. RF models multiple decision trees to generate aggregated predictions, which improve accuracy while preventing model overfitting. However, unlike Random Forest, ETR introduces additional randomness by selecting split thresholds thoroughly at random, rather than optimizing them. This randomness makes the trees more diverse and often reduces variance, leading to better generalization in some cases. The final prediction in ETR is obtained by averaging the outputs of all the individual trees. A key advantage of ETR is its efficiency, as the random split selection reduces computational cost compared to fully optimized splits. It is particularly effective for high-dimensional datasets and complex non-linear relationships, as it captures interactions between variables without requiring explicit feature engineering [25].

2.2.2. *Quantile Regression (QR)*

QR functions as a statistical and ML method which enhances standard regression by predicting conditional quantiles of the response variable, including median and 25th/75th percentiles, instead of calculating average values. Unlike ordinary least squares (OLS), which minimizes the

Table 1. VIF feature selection.

	Initial VIF score	Final VIF score
Energy Consumption (kWh)	171.985	Removed
Temperature ($^{\circ}$C)	107.168	1.631
Humidity (%)	220.049	Removed
Occupancy Rate (%)	87.506	1.428
Lighting Consumption (kWh)	119.025	Removed
HVAC Consumption (kWh)	131.099	Removed
Energy Price ($/kWh)	246.694	Removed
Carbon Emission Rate (gCO_2/kWh)	194.312	Removed
Power Factor	1730.818	Removed
Voltage Levels (V)	5047.496	Removed
Reactive Power (kVARh)	172.526	Removed
Indoor Temperature ($^{\circ}$C)	466.545	Removed
Building Age (years)	109.181	5.737
Equipment Age (years)	116.523	Removed
Energy Efficiency Rating	316.140	Removed
Building Size (m^2)	112.460	Removed
Window-to-Wall Ratio (%)	239.429	Removed
Insulation Quality Score	323.026	Removed
Historical Energy Consumption (kWh)	136.509	Removed
Local Energy Production (kWh)	81.591	6.557
Grid Stability Score	1434.947	Removed
Solar Irradiance (W/m^2)	118.390	Removed
Smart Plug Usage (kWh)	105.300	8.091
Water Usage (liters)	113.252	Removed
Energy Savings Target (%)	235.108	Removed
Room-Level Energy Consumption (kWh)	118.165	Removed
Zonal Heating/Cooling Data (kWh)	114.409	Removed
IoT Sensor Count	162.208	Removed
Thermal Comfort Index	555.771	Removed

squared errors to predict the average outcome, QR minimizes a weighted absolute loss function, allowing it to model the whole conditional distribution of the dependent variable. This makes QR particularly powerful when the data exhibit heteroscedasticity (non-constant variance) or when extreme values and distribution tails are of interest. For example, QR can be used to study how predictors influence not only the average response but also the variability and potential risks (upper or lower quantiles) [26].

2.2.3. *Stochastic Gradient Boosting (SGB)*

The SGB is an advanced boosting learning method that builds upon the concept of gradient boosting by introducing randomness to enhance generalization and mitigate overfitting. In gradient boosting, trees are built one after another, with each new tree successively correcting the residual errors using gradient descent optimization. SGB enhances this process by randomly sampling a subset of the training data (row sampling) and/or features (column sampling) at each iteration, a technique similar to bagging. The stochastic element reduces variance, prevents overfitting, and often improves predictive performance, especially in large or noisy datasets. The SGB algorithm provides maximum adaptability because it supports the optimization of various loss functions, including MSE, quantile loss, and classification loss. It works for both regression and classification problems. One of the major reasons behind its widespread use in various real-life scenarios is its ability to handle non-linear relationships, intricate feature interactions, and its robustness against outliers. Its application ranges from energy forecasting to financial markets as well as structural reliability studies [27].

2.2.4. *Kepler Optimization Algorithm (KOA)*

KOA is a nature-inspired metaheuristic optimization technique that effectively captures the essence of Kepler's laws of planetary motion through clever metaphorical modeling. It essentially represents the planetary movement around the sun, taking into consideration the orbital velocity, the gravitational pull, and the elliptical revolutions of the planet, leading the search process to a solution that is the best possible. Those candidate solutions contained in KOA are brought to the forefront as planets whose movement in the exploration space is influenced by rules based on Kepler's laws. As a result, there is a perfect mix of exploration (searching for global optima in a wide manner)

and exploitation (utilizing resources more closely around a promising solution). KOA can handle complex problems that are non-linear and multi-dimensional, particularly when it excels in escaping local minima, and is relatively easy to implement. Engineering design, energy systems optimization, ML feature selection, and scheduling all show the presence of these signatures. The combination of physical inspiration with computational efficiency makes it effective [28].

2.2.5. *Gazelle Optimization Algorithm (GOA)*

GOA is a new metaheuristic optimization algorithm prompted by the social and survival behaviors of gazelles in the wild. The algorithm simulates how gazelles find food, maintain group coordination, and respond to predators. On a high level, the algorithm represents both exploration (looking for new areas of the solution space, which is similar to gazelles dispersing to find food) and exploitation (deepening the search around promising areas, like gazelles moving towards a few resources). Also, the predator-prey relationship has been conceptualized to facilitate the algorithm in overcoming local minima and improving the performance of global optimization. GOA has been identified as having significant potential in dealing with difficult, high-dimensional, and non-linear problems due to its flexibility and the balance between exploration and exploitation. It has been applied successfully in engineering, ML, feature selection, and energy systems. In comparison with traditional algorithms, such as Genetic Algorithms (GA) or Particle Swarm Optimization (PSO), GOA can be more effective in cases where it is capable of reaching the goal in less time, and thus showing a higher degree of resistance to being trapped in local optima [29].

2.3. Evaluation Metrics

In this research, the number of evaluation metrics is varied to ensure that various dimensions of the performance of regression models are captured. The R2, RMSE, MAE, and MARD are metrics reflecting various types of error level and accuracy, as well as COV is a measure of the stability of the predictions, and PI is a measure of how much of the uncertainty is captured by the model. They combine to create a multi-angle and balanced perspective of predictive reliability. In order to make a formal statement of these measures, let A_i denote the actual values, P_i the predicted values, $\bar{A}$ the mean of actual values, N the sample size, σ the standard deviation of predictions, μ the mean of predictions, and $t_{\alpha/2,(N-2)}$ the critical t value for the prediction interval. The evaluation metrics are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(A_i - P_i)^2}{\sum_{i=1}^{N}(A_i - \bar{A})^2} \quad (1)$$

$$\text{MARD} = \text{Median}\left(\frac{|A_i - P_i|}{|A_i|}\right) \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2} \quad (3)$$

$$\text{COV} = \frac{\sigma}{\mu} \quad (4)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|A_i - P_i| \quad (5)$$

$$PI = \bar{A}^2 \pm t_{\alpha/2,(N-2)} \times \quad (6)$$

2.4. K-fold Cross Validation

Cross-validation is a resampling method that helps to measure the effectiveness and the applicability of ML models. Several subsets (or "folds") are created from the dataset, and the model is taught using some folds and tested on the remaining ones; thus, every fold has been used as a test set in turn. To attain an accurate assessment of the model's correctness and reduce the likelihood of overfitting, results are combined. A typical example of such methods is k-fold cross-validation, where data is classified into k equal-sized folds. The method allows researchers to assess model performance through different data segments, which leads to better evaluation stability and fairness in the results [30].

Fig. 1 highlights the outcomes of the 5-fold cross-validation for the three predictive models, namely QR, ETR, and SGB, along with their corresponding RMSE and R^2 values. With respect to the QR model, the best fold was K5, which had the lowest RMSE (0.3103) and the highest R^2 (0.9168). Concerning ETR, fold K5 was simultaneously the most robust and the most divergent from the rest, as it presented the lowest RMSE (0.2837) and the highest R^2 (0.9257), thus making the generalization better than the rest of the folds. For SGB as well, the best-performing fold was K5, which had the lowest RMSE (0.2713) and the highest R^2 (0.9359), thereby exceeding all other folds in terms of both accuracy and predictive strength. Overall, while all models showed consistent stability across folds, the best-performing folds for QR, ETR, and SGB were consistently identified as K5, with SGB delivering the most accurate and robust predictions among the three models.

3. Results and discussion

3.1. Hyperparameter Optimization and Convergence Analysis

Hyperparameter tuning is the process of systematically determining the optimal values for parameters that control the learning process of an ML model (e.g., learning rate, number of estimators, or regularization strength). Since

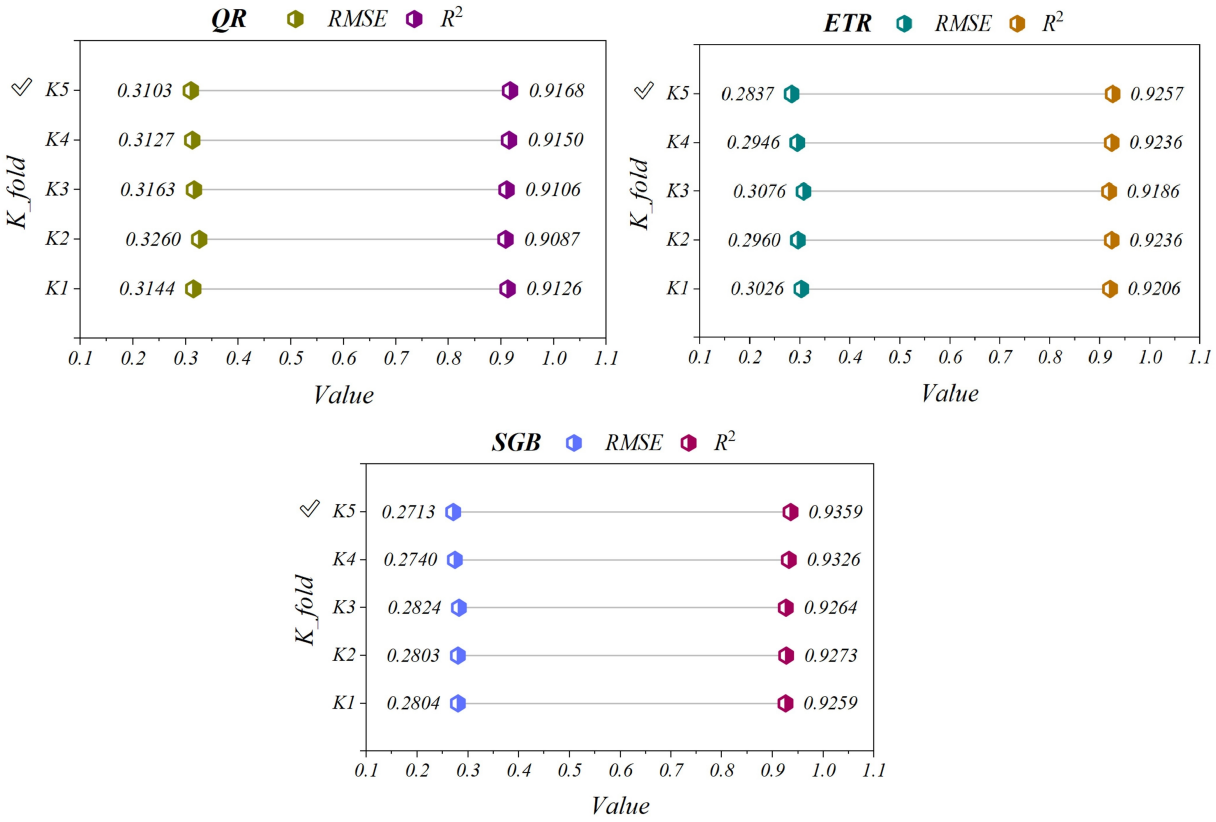

Fig. 1. Results of 5-fold cross-validation for the baseline models.

these parameters are not learned directly from the data, tuning is performed using search strategies such as grid search, random search, or metaheuristic optimization to maximize model performance while avoiding overfitting or underfitting [31]. Convergence analysis evaluates whether the training process of a model or optimization algorithm stabilizes over iterations, typically by monitoring loss functions or performance metrics until they plateau. It ensures that the algorithm reaches an optimal or near-optimal solution without unnecessary computations or divergence. Together, hyperparameter tuning and convergence analysis improve both the efficiency and reliability of model training [32].

3.2. Results of the Predictive Models Based on Metrics

Table 2 presents the prediction proficiency of the studied models in the test phase, providing a clear comparison between the baseline models (QR, ETR, SGB), their GA-optimized versions (QRGA, ETGA, SGGA), and the KOA-optimized counterparts (QRKA, ETKA, SGKA). Overall, the results highlight the strong advantage of optimization, particularly with KOA. The baseline QR performed the weakest, with an R^2 of 0.909 and the highest RMSE of 0.320, indicating less accurate predictions. Introducing GA optimization improved performance moderately (QRGA, R^2 = 0.937, RMSE = 0.265), while KOA optimization provided a more substantial gain (QRKA, R^2 = 0.945, RMSE = 0.244). A similar trend was observed for ETR, where the baseline achieved R^2 = 0.936 with RMSE = 0.267; however, ETKA outperformed it significantly, with R^2 = 0.960 and RMSE = 0.212. SGB showed comparable baseline accuracy (R^2 = 0.934, RMSE = 0.266). However, its KOA-enhanced version, SGKA, achieved the best overall performance, with R^2 = 0.975 and the lowest RMSE = 0.162, as well as the lowest MAE and MARD values, indicating superior accuracy and reliability. These results indicate practical improvements in predictive accuracy with KOA optimization across models, with SGKA showing the best numerical performance (highest R^2 and lowest RMSE, MAE, and MARD) for estimating energy-saving potential in the test phase.

Fig. 2 illustrates scatter plots and error histograms of prediction points. The first set of charts is residual plots, which show the differences between predicted and actual values (residuals) against predicted outputs, helping to evaluate model bias and variance across training, validation, and test sets. The second set of charts is error distribution his-

Table 2. Results of performance evaluation of prediction models on train, validation, and test sets.

		R^2	RMSE	COV	PI (%)	MAE	MARD
Train (n = 1753, 80%)	QR	0.918	0.305	0.027	92.698	0.222	0.019
	QRGA	0.935	0.272	0.024	92.869	0.197	0.017
	QRKA	0.945	0.249	0.022	93.782	0.181	0.016
	ETR	0.929	0.282	0.025	93.212	0.203	0.018
	ETGA	0.942	0.252	0.023	93.953	0.184	0.016
	ETKA	0.960	0.211	0.019	92.755	0.152	0.013
	SGB	0.931	0.275	0.025	92.698	0.200	0.017
	SGGA	0.962	0.203	0.018	93.668	0.149	0.013
	SGKA	0.974	0.167	0.015	93.155	0.122	0.011
Validation (n=219, 10%)	QR	0.897	0.343	0.030	94.064	0.255	0.022
	QRGA	0.918	0.296	0.026	94.064	0.219	0.019
	QRKA	0.939	0.250	0.022	94.064	0.187	0.016
	ETR	0.923	0.280	0.025	93.607	0.209	0.018
	ETGA	0.935	0.272	0.024	94.977	0.206	0.018
	ETKA	0.950	0.228	0.020	93.607	0.172	0.015
	SGB	0.920	0.305	0.027	93.151	0.227	0.020
	SGGA	0.951	0.223	0.020	94.977	0.175	0.015
	SGKA	0.971	0.175	0.015	95.434	0.136	0.012
Test (n=219, 10%)	QR	0.909	0.320	0.028	94.521	0.241	0.021
	QRGA	0.937	0.265	0.024	94.521	0.199	0.017
	QRKA	0.945	0.244	0.022	92.237	0.183	0.016
	ETR	0.936	0.267	0.024	95.890	0.198	0.017
	ETGA	0.946	0.240	0.021	94.064	0.176	0.015
	ETKA	0.960	0.212	0.019	93.151	0.158	0.014
	SGB	0.934	0.266	0.023	93.151	0.197	0.017
	SGGA	0.960	0.211	0.019	94.064	0.154	0.013
	SGKA	0.975	0.162	0.014	93.151	0.118	0.010

tograms, which present the frequency of percentage errors, indicating how concentrated or dispersed the prediction errors are around zero. For a better understanding, the productivity of the baseline QR model is reviewed and juxtaposed with that of its hybrid counterpart, QRKA. In the QR model, the residuals exhibit wider scatter, especially at higher predicted values, while the error histogram, although centered around zero, shows a broader spread, suggesting higher variability in predictions.

In contrast, QRKA residuals are more tightly clustered around zero, with a narrower and more peaked error distribution, reflecting improved accuracy and stability. This demonstrates that QRKA outperformed QR by reducing bias and variability. It is worth noting that this performance trend—where the hybrid models consistently achieve better prediction accuracy and error concentration than their single-model counterparts—was observed across all model comparisons in the study, highlighting the overall benefit of hybridization. Fig. 4 shows box plots merged with violin-style error distributions to provide a more comprehensive visualization of the spread and density of prediction errors for different models. The box plots also indicate the median error, the interquartile range, as well as the total variability. The dots along the distribution depict the closeness of the error values to zero by their density. As can be seen from the distribution of errors for the ETR, the errors are more or less equally distributed; however, the range is considerably larger compared to the combined version. The ETKA model, on the other hand, exhibits a more compact error range with a denser clustering of points near zero, indicating reduced variability and improved predictive consistency. This comparison suggests that while both models perform well, ETKA delivers more stable and accurate predictions with fewer extreme deviations, highlighting the advantage of the hybrid optimization in refining the baseline ETR model. Table 3 displays the results of the Wilcoxon signed-rank test, which was applied to statistically compare the prediction errors of the models. The p-values across all models, including both the baseline (QR, ETR, SGB) and their hybrid or optimized versions (QRGA, QRKA, ETGA, ETKA, SGGA, SGKA), are all greater than 0.05. This indicates that there are no notable variations between the prediction errors of the models when compared under the null hypothesis. In other words, while the hybrid and optimized models showed performance improvements in terms of accuracy and error distribution, these differences were not statistically significant at the 95% confidence level. The results, therefore, suggest that the

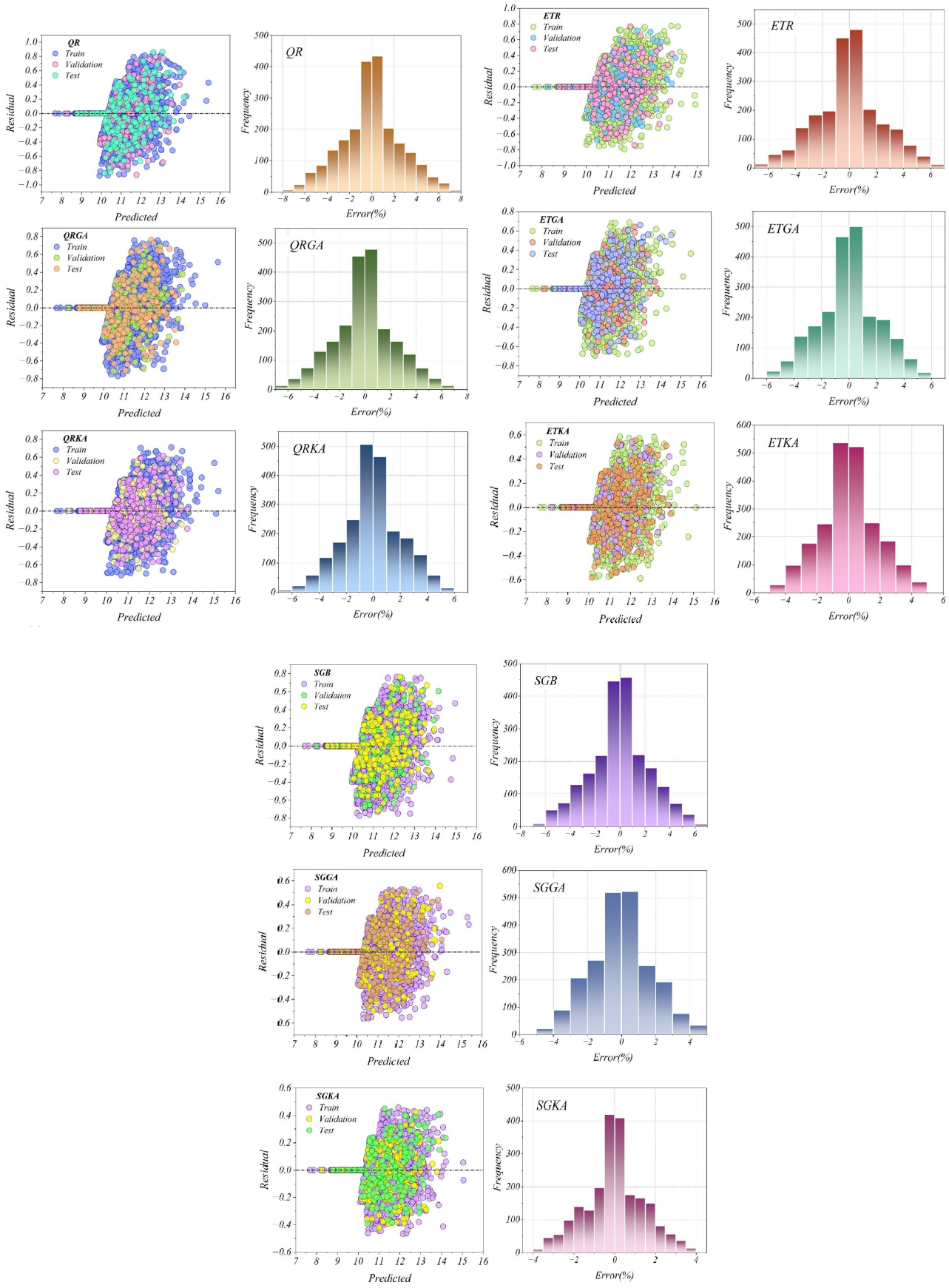

Fig. 2. Scatter plots of predicted versus actual energy savings potential and histograms of percentage prediction errors for the QR and QRKA models on the test set (n=219).

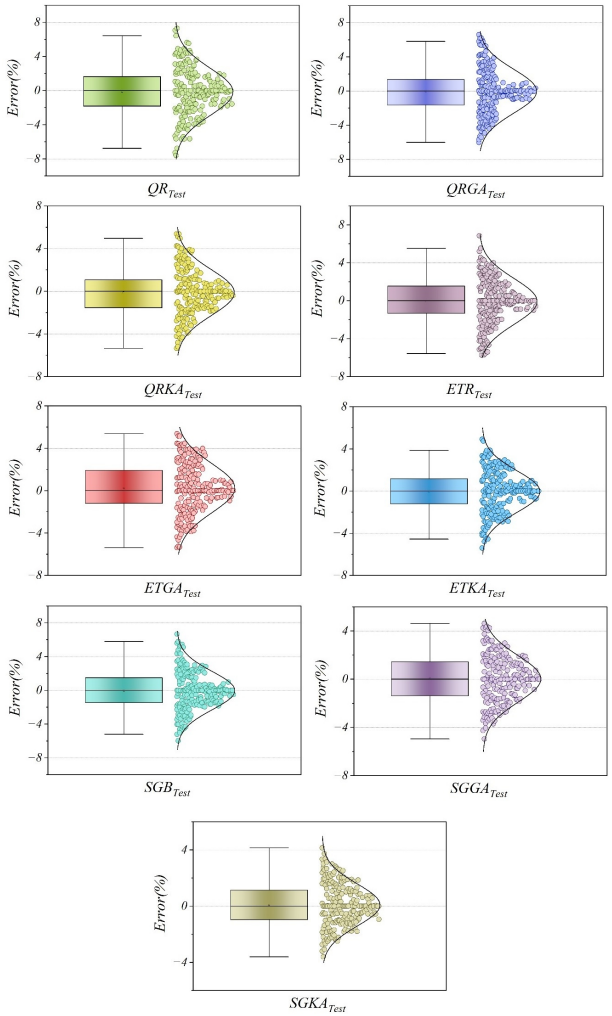

Fig. 3. Scatter plots of predicted versus actual energy savings potential and histograms of percentage prediction errors for the QR and QRKA models on the test set (n=219).

observed performance gains, although practically meaningful, may not be statistically significant enough to reject the hypothesis of similarity in model performance.

3.3. Feature Importance Analysis Using SHAP

Although the Wilcoxon signed-rank test indicated no statistically significant differences in prediction errors among the models, it is important to understand which input variables drive the predictions of the best-performing model (SGKA). To this end, SHAP values were computed for the SGKA model on the test set. SHAP is a unified framework based on cooperative game theory that assigns each feature an importance value for a particular prediction. It provides both global feature rankings and local explanations, making it particularly suitable for interpreting complex nonlinear ensemble models such as Stochastic Gradient Boosting.

Fig. 4 presents the SHAP summary plot (beeswarm) for the SGKA model. Features are ranked vertically by their mean absolute SHAP value, reflecting their overall importance in driving energy-saving potential predictions. Each dot represents a single observation (test sample), with the horizontal position indicating the SHAP value's impact on the model output (positive values increase the predicted energy-saving potential, while negative values decrease it). The color scale represents the feature value. The plot reveals that Energy Consumption (kWh) is by far the most influential feature, followed by Temperature (°C) and Occupancy Rate (%). Higher actual energy consumption strongly increases the predicted savings potential, which is intuitively consistent with the definition of the target variable. Temperature and occupancy rate also show substantial and nonlinear effects, with higher temperatures and occupancy levels generally contributing positively or negatively depending on interactions with other variables. Other notable features include Local Energy Production (kWh), Building Age (years), and Smart Plug Usage (kWh), which exhibit moderate but meaningful contributions. Features with lower importance (e.g., Thermal Comfort Index) have narrower spreads and smaller impacts.

This SHAP analysis demonstrates that the SGKA model effectively captures complex, nonlinear interactions among the retained variables. The insights gained can inform practical retrofit decisions by highlighting the key drivers of energy-saving potential for the type of system studied.

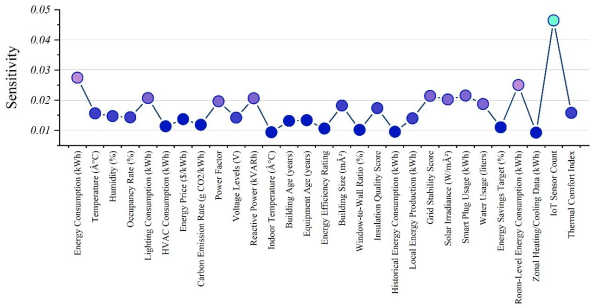

Fig. 4. SHAP summary plot for the best-performing model.

4. Conclusion

In this study, predictive models were developed and evaluated to estimate energy-saving potential using a dataset of 2,191 aggregated daily samples derived from 52,584 raw measurements averaged over 24-hour intervals. A total of

Table 3. Results of the Wilcoxon test.

Model					
QR		ETR		SGD	
p_ value	Statistic	p_value	Statistic	p_ value	Statistic
0.54178	1173947.5	0.5350	1168256	0.715443147	1185522
QRGA		ETGA		SGGA	
p_ value	Statistic	p_ value	Statistic	p_value	Statistic
0.5197	1171877	0.6060	1176725.5	0.3330	1167698.5
QRKA		ETKA		SGKA	
p_value	Statistic	p_value	Statistic	p_value	Statistic
0.4208	1171435	0.6092	1175780	0.9463	1188845

29 input variables and one output variable representing energy-saving potential were considered.

- Initially, feature selection has been conducted with the help of the Variance Inflation Factor (VIF) procedure. The threshold applied was 10 to eliminate multicollinearity.
- In the next step, the researchers have used three ML models —Extra Trees Regression (ETR), Quantile Regression (QR), and Stochastic Gradient Boosting (SGB).
- Their work was further improved through hyperparameter tuning by the Kepler Optimization Algorithm (KOA) and the Gazelle Optimization Algorithm (GOA).
- The evaluation of the models was done through 5-Fold Cross-Validation.
- The prediction accuracies and the robustness of the models were measured with R^2, RMSE, COV, PI, MAE, and MARD metrics, while run time was also included.
- The Wilcoxon test was used to verify the statistical robustness, and SHAP analysis was performed to identify the most influential features and their impact directions.
- In general, the research has revealed possible efficient ways to achieve increased accuracy, stability, and interpretability of energy-saving potential predictions. Results revealed that out of all the models, the KOA-optimized SGB (SGKA) was the most accurate in the test phase with an R^2 of 0.975 and the lowest RMSE of 0.162, as well as the smallest MAE and MARD values, confirming that it was the most reliable in terms of predictions. Firstly, the pictures of residuals and the histograms of error have also been presented as evidence that the hybrid models, in general, were able to decrease both the bias and the variability compared to their baseline counterparts.
- Secondly, the comparison between measured and predicted values has demonstrated that all models exhibit good statistical characteristics in relation to the actual data. In particular, the optimized hybrids (e.g., QRKA, ETKA, SGKA) have been successful in achieving better stability and a closer match with the measured values.
- SHAP analysis revealed that Energy Consumption (kWh), Temperature (°C), and Occupancy Rate (%) are the most influential features, while the overall improvements in predictive accuracy arise primarily from the model's ability to capture complex nonlinear interactions among the retained variables.

Overall, the findings confirm the practicality and robustness of the proposed hybrid modeling framework, demonstrating that combining ML models with advanced optimization techniques can provide reliable, accurate, and interpretable predictions of energy-saving potential.

References

[1] A. Shahsavari and M. Akbari, (2018) *"Potential of solar energy in developing countries for reducing energy-related emissions"* **Renewable and sustainable energy reviews *90***: 275–291. DOI: 10.1016/j.rser.2018.03.065.

[2] M. M. Bah and M. Y. Saari, (2020) *"Quantifying the impacts of energy price reform on living expenses in Saudi Arabia"* **Energy Policy *139***: 111352. DOI: 10.1016/j.enpol.2020.111352.

[3] A. Allouhi, Y. E. Fouih, T. Kousksou, A. Jamil, Y. Zeraouli, and Y. Mourad, (2015) *"Energy consumption and efficiency in buildings: current status and future trends"* **Journal of Cleaner production *109***: 118–130. DOI: 10.1016/j.jclepro.2015.05.139.

[4] A. Martos, R. Pacheco-Torres, J. Ordóñez, and E. Jadraque-Gago, (2016) *"Towards successful environmental performance of sustainable cities: Intervening sectors. A review"* **Renewable and Sustainable Energy Reviews *57***: 479–495. DOI: 10.1016/j.rser.2015.12.095.

[5] B. Lin and J. Zhu, (2019) *"Impact of energy saving and emission reduction policy on urban sustainable development: Empirical evidence from China"* **Applied Energy *239***: 12–22. DOI: 10.1016/j.apenergy.2019.01.166.

[6] S. A. H. Zaidi, M. Hussain, and Q. U. Zaman, (2021) *"Dynamic linkages between financial inclusion and carbon emissions: evidence from selected OECD countries"* **Resources, Environment and Sustainability *4***: 100022. DOI: 10.1016/j.resenv.2021.100022.

[7] K. S. Moon, (2008) *"Sustainable structural engineering strategies for tall buildings"* **The Structural Design of Tall and Special Buildings *17***: 895–914. DOI: 10.1002/tal.475.

[8] J.-K. Choi, J. Eom, and E. McClory, (2018) *"Economic and environmental impacts of local utility-delivered industrial energy-efficiency rebate programs"* **Energy Policy *123***: 289–298. DOI: 10.1016/j.enpol.2018.08.066.

[9] V. Jadhav, R. Jadhav, P. Magar, S. Kharat, and S. U. Bagwan. "Energy conservation through energy audit". In: *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. IEEE, 2017, 481–485. DOI: 10.1109/ICOEI.2017.8300974.

[10] I. Johansson, S. Johnsson, and P. Thollander, (2022) *"Impact evaluation of an energy efficiency network policy programme for industrial SMEs in Sweden"* **Resources, Environment and Sustainability *9***: 100065. DOI: 10.1016/j.resenv.2022.100065.

[11] F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli, and J. N. Kutz, (2019) *"Data-driven modeling and learning in science and engineering"* **Comptes Rendus Mécanique *347***: 845–855. DOI: 10.1016/j.crme.2019.11.009.

[12] X. Yang, S. Liu, Y. Zou, W. Ji, Q. Zhang, A. Ahmed, X. Han, Y. Shen, and S. Zhang, (2022) *"Energy-saving potential prediction models for large-scale building: A state-of-the-art review"* **Renewable and Sustainable Energy Reviews *156***: 111992. DOI: 10.1016/j.rser.2021.111992.

[13] P. Shen, Z. Wang, and Y. Ji, (2021) *"Exploring potential for residential energy saving in New York using developed lightweight prototypical building models based on survey data in the past decades"* **Sustainable Cities and Society *66***: 102659. DOI: 10.1016/j.scs.2020.102659.

[14] J. Ko, J. Park, and J.-W. Jeong, (2021) *"Energy saving potential of a model-predicted frost prevention method for energy recovery ventilators"* **Applied Thermal Engineering *185***: 116450. DOI: 10.1016/j.applthermaleng.2020.116450.

[15] Y. Xu, J. Winter, and W.-C. Lee. "Prediction-based strategies for energy saving in object tracking sensor networks". In: *IEEE International Conference on Mobile Data Management, 2004. Proceedings. 2004.* IEEE, 2004, 346–357. DOI: 10.1109/MDM.2004.1263084.

[16] T. M. Chung and J. Burnett, (2001) *"On the prediction of lighting energy savings achieved by occupancy sensors"* **Energy engineering *98***: 6–23. DOI: 10.1080/01998590109509317.

[17] W. Kleiminger, S. Santini, and F. Mattern. *Simulating the energy savings potential in domestic heating scenarios in Switzerland*. 2014. DOI: 10.3929/ethz-a-010193004.

[18] H.-I. Cho, D. Cabrera, and M. K. Patel, (2020) *"Estimation of energy savings potential through hydraulic balancing of heating systems in buildings"* **Journal of Building Engineering *28***: 101030. DOI: 10.1016/j.jobe.2019.101030.

[19] K. B. Wittchen, J. Kragh, and O. M. Jensen, (2011) *"Energy saving potentials–A case study on the Danish building stock"* **SBi, Aalborg University**:

[20] M. Y. M. Yusop. *Energy saving for pneumatic actuation using dynamic model prediction*. Cardiff University (United Kingdom), 2006.

[21] E. Widayati, (2021) *"Estimating Energy Saving Potential by Taking Into Account Interdependence Effect Case Study: High Rise Office Building in Jakarta"* **Jurnal Inovasi Pendidikan dan Sains 2**: 26–32. DOI: 10.51673/jips.v2i1.476.

[22] C. Zhang, F. Zhao, and Y. Liu, (2012) *"Diagrid tube structures composed of straight diagonals with gradually varying angles"* **The Structural Design of Tall and Special Buildings *21***: 283–295. DOI: 10.1002/tal.596.

[23] D. Engineer. *Southern California Energy Consumption Dataset*. Kaggle, 2024.

[24] R. M. O'brien, (2007) *"A caution regarding rules of thumb for variance inflation factors"* **Quality & quantity *41***: 673–690. DOI: 10.1007/s11135-006-9018-6.

[25] P. Geurts, D. Ernst, and L. Wehenkel, (2006) *"Extremely randomized trees"* **Machine learning *63***: 3–42. DOI: 10.1007/s10994-006-6226-1.

[26] R. Koenker and G. B. Jr, (1978) *"Regression quantiles"* **Econometrica: journal of the Econometric Society**: 33–50. DOI: 10.2307/1913643.

[27] J. H. Friedman, (2002) *"Stochastic gradient boosting"* **Computational statistics & data analysis *38***: 367–378. DOI: 10.1016/S0167-9473(01)00065-2.

[28] S. H. Hakmi, A. M. Shaheen, H. Alnami, G. Moustafa, and A. Ginidi, (2023) *"Kepler algorithm for large-scale systems of economic dispatch with heat optimization"* **Biomimetics *8***: 608. DOI: 10 . 3390 / biomimetics8080608.

[29] J. O. Agushaka, A. E. Ezugwu, and L. Abualigah, (2023) *"Gazelle optimization algorithm: a novel nature-inspired metaheuristic optimizer"* **Neural Computing and Applications *35***: 4099–4131. DOI: 10 . 1007 / s00521-022-07854-6.

[30] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. ***14***. Montreal, Canada, 1995, 1137–1145.

[31] J. Bergstra and Y. Bengio, (2012) *"Random search for hyper-parameter optimization."* **Journal of machine learning research *13***:

[32] L. Bottou, F. E. Curtis, and J. Nocedal, (2018) *"Optimization methods for large-scale machine learning"* **SIAM review *60***: 223–311. DOI: 10 . 1137 / 16M1080173.