SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 9528-9539
Open Access. Online & Print

DOI: 10.5281/zenodo.20158499

STATISTICAL MODELING AND OPTIMIZATION OF ENGINEERING SYSTEMS USING ADVANCED DATA ANALYTICS TECHNIQUES

[1]Uma N, [2]Dr. Ketul Brahmbhatt, [3]Neeta Chudasama, [4]Dr. Pankaj Agarwal, [5]Dr. Ch.Raja, [6]Mr. Ankush Madhukar Gund, [7]A. Vijayakumar

[1]Assistant Professor, Department of Information Technology, Sri Venkateswara College of Engineering , 602117, Chennai, India, Email Id : numa@svce.ac.in, Orcid Id : 0000-0001-5242-6845
[2]Assistant Professor, Department of Mechanical Engineering, Birla Vishvakarma Mahavidhyalaya Vallabh, Vidhyanagar , Gujarat Email Id : kbrahmbhatt@bvmengineering.ac.in
[3]Assistant Professor, Department of Computer Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Faculty of Technology & Engineering(FTE), Charotar University of Science & Technology (CHARUSAT),Changa, Gujarat, India Email Id : neeta.chudasama2011@gmail.com
[4]Professor and Dean, Specialization in Data science and ML, School of Engineering and Technology 201017, Ghaziabad, Email Id : pankaj7877@gmail.com, Orcid Id : 0000-0002-1027-2419
[5]Associate Professor, Department of ECE, Specialization in Signal processing ECE, Mahatma Gandhi Institute of:500075 , Hyderabad , India , Email Id : chraja@mgit.ac.in
[6]Assistant Professor, Department of Instrumentation,Specialization in Automation,IOT and Embedded system , Bharati Vidyapeeth College Of Engineering, Navi Mumbai, Maharashtra , 400614 , India, Email Id : ankush.gund@bvcoenm.edu.in, Orcid Id: 0000-0002-1294-2850
[7]Professor, Department of Civil Engineering, V.S.B. Engineering College, Karur, Tamil Nadu, India – 639111, Email Id : vijayakumarkct@gmail.com

Received: *15/10/2025*
Accepted: *15/01/2026*

Corresponding Author: Uma N (numa@svce.ac.in)

Abstract

A statistical modeling and optimization framework of engineering systems is introduced to respond to the increasing demand of sustainable, data-driven decision making in the energy-intensive applications. As an example, a residential building energy system is considered, and heating and cooling loads are modeled as a function of the important geometric and envelope-related design variables based on a simulation-based dataset. The suggested model involves Response Surface Methodology, Generalized Additive Models and Random Forest regression to address the linear, nonlinear and interaction effect and after that systematic analysis of the model performance in terms of various measures of accuracy is carried out. The predictive models that are most accurate are then incorporated in a multi-objective optimization process using the NSGA-II algorithm in order to derive Pareto-optimal building design configurations. The findings indicate that machine learning models are more predictive accurate than classical statistical methods in particular with regard to cooling load prediction whereas statistical models maintain high levels of interpretability and explanatory power. The optimization analysis shows that the trade-off between heating and cooling goals is rather small, which means that energy-efficient design methods can be used to increase both of the measures. Comprehensively, the results prove that statistical analysis, sophisticated data analytics, and evolutionary optimization can be a strong and efficient framework in terms of engineering system analysis and optimization.

Keywords: *Statistical modeling; Advanced data analytics; Machine learning; Multi-objective optimization; Building energy efficiency*

1. Introduction

Modern technology development is based on engineering systems which contribute to very important work in the energy industry, manufacturing and transportation, and the industrial infrastructure. These systems are becoming more sensitive to various interacting design parameters, and their performance is clustered as the size and complexity of these systems increases. Such complex behavior is no longer possible to be fully represented by traditional methods of engineering analysis, which tend to make simplified assumptions or investigate parameters separately. This has led to an increasing trend of data-driven methods who integrate statistical modelling, data analytics and optimization algorithm to facilitate informed decision-making in engineering design and operation [1].

The recent years have witnessed the fast rising use of analytics-based approaches due to the accessibility to large-scale engineering data and the growth of computational power. Statistical modeling is still a fundamental instrument of relations between inputs and outputs of systems, with interpretability and theoretical rigor. Meanwhile, high-quality predictive performance using more sophisticated data analytics, such as machine learning, has been shown to be better in nonlinear interactions and higher-dimensional design spaces. The combination of these methods allows the engineers to get out of the descriptive analysis to the predictive as well as prescriptive optimization of the system [2].

High-level data analytics has already been effectively implemented in various areas of engineering such as optimization of processes, quality assurance, and performance improvement of the system. Using historical and simulation generated data, engineers may build surrogate models that are very accurate approximations of the complexity of a system. These models minimize the cost of experimentation or computationally expensive simulations but allow the quick assessment of alternative design situations. Typically, these capabilities are useful in engineering systems where performance targets are competing or limited by physical and working conditions [3].

Statistical modeling (i.e. regression analysis and response surface methodology) is also very important in this scenario as it offers structural representations of the system behavior. These approaches permit the explicit quantification of the main effects, and the effect of interaction, and the curvature in the response surface, which is why they are highly applicable in the engineering context, where interpretability and transparency are vital. Engineering statistics is still used to maintain the connection between theoretical modeling and practical decision making so that information provided by data is always based on solid analytical underpinnings [4].

In addition to the conventional statistical methods, the new trend of advanced analytics has also ventured into areas that have historically been linked to management and operational decision-making. Although most of such studies are dealt with the business or operational context, the same principles of analysis can be applied to engineering systems, especially those with complex trade-offs and limited design spaces [5].

Analytics-based optimization has been identified to enhance efficiency, productivity and resource utilization of different systems.

Although the advantages of using analytics-based optimization have been demonstrated, there are still a number of shortcomings in the engineering literature. Numerous studies are dedicated to single modeling methods with no systematic comparison between statistical and machine learning methods. Others lay stress on predictive accuracy without properly suggesting how the predictive models can be incorporated within optimization models. Further, the interpretability of the advanced analytics models is not always considered, and this constrained their use in engineering where transparency and explainability are essential [6].

One of the key problems in the engineering practice is optimization, especially in the situation when a system is controlled by several competing goals. Such problems have been long solved using mathematical optimization methods, but these methods are susceptible to the unavailability of explicit objective functions and gradients. On the other hand, recent optimization algorithms, such as evolutionary and population-based algorithms, are better adapted to data-driven environments, in which the objective functions are estimated by surrogate models. These methods allow an effective search of nonlinear design spaces as well as dealing with nonlinearities and constraints of real-world engineering systems [7].

The combination of machine learning and optimization has also increased the promise of

data-driven engineering. Machine learning models can be used as precise surrogates to complex systems, and optimization algorithms can then be used to find high-performing design configurations by using these surrogates. This integration has been successful in areas like the supply chain management and logistics where the performance goals are to be balanced with the sustainability and the operational constraints. These advancements indicate the wider use of analytics-based optimization in engineering, which is less specific to engineering borders [8].

One of the most critical areas of application of advanced data analytics and optimization are energy systems. Over the years, the growing energy demand in the world coupled with the sustainability requirements has enhanced the necessity of developing and operating efficient energy systems. Energy forecasting, performance evaluation, and optimization of the system have seen extensive usage of predictive analytics and machine learning to offer data-driven solutions to the reduction of consumption and enhancement of efficiency. Such strategies are nowadays considered as critical instruments towards attainability of sustainable energy infrastructure [9].

In the energy world, construction of energy systems are the largest sources of total energy consumption. Numerous design variables affect heating and cooling loads such as geometry, material properties, orientation, and glazing properties. These relationships are difficult to model well as they are nonlinear, and there are performance goals competing with each other. The solution to this complexity is offered by advanced analytics, which captures the delicate patterns in data and helps to proceed the optimization-based design strategies, which trade off among various energy objectives [10].

Still, the current literature usually covers either prediction or optimization separately and does not develop an integrated framework to bring together statistical modeling, advanced analytics, and multi-objective optimization. Such disaggregations restrict the usefulness of research results especially in early-stage design decision-making that requires consideration of trade-offs on a holistic basis. It is evident that there is a requirement of integrated methodologies that are characterized by interpretability, predictive accuracy, and optimization capability in a single analytical framework [11].

This research fulfills this requirement by providing a combined framework of statistical modeling and optimization of engineering systems by using applied data analytics. As an example of the engineering application, the building energy performance is considered, and the statistical and machine learning models to predict heating and cooling loads are evaluated systematically and the most effective ones integrated into the multi-objective optimization algorithm. The research is limited in terms of the variables of the study design and energy performance indicators, which is methodologically clear and repeatable. Although external influences, including occupant behavior and economic aspects that cannot be covered within the framework of this work, the suggested framework can be generalized and applied to other engineering systems and optimization goals.

This research is important in that it contributes to the data-driven engineering methodology. The paper combines several data analytical paradigms in one framework to show how sophisticated data analytics can improve the cognition as well as optimization of sophisticated engineering systems. The results are conducive to the production of eco-friendly, effective, and smart engineering solutions that are in line with the new technological trends. To this end, the research questions are as follows: (1) to design and compare statistical and machine learning models to predict the performance of engineering systems; (2) to determine the appropriateness of the models to be incorporated into multi-objective optimization formats; and (3) to determine how to work out the best system configurations that trade competing performance goals through the use of data-driven optimization.

2. Literature Review

The use of data-driven methods in energy and engineering systems has grown accelerated in the last decade due to the developments in the field of artificial intelligence, machine learning, and computational optimization. Within the energy systems, data science has turned out to be a significant instrument in consumption pattern knowledge, finding efficiency prospects, and guiding informed choices. In a more detailed overview, Ohalete et al. note the increasingly popular application of artificial intelligence and data analytics to be able to analyze the behavior of energy consumption in general, and indicate the intricate relationships between the variables of the system and its energy performance, which are challenging to describe using the traditional analytical tools [12]. In their work, they highlight the increased significance of predictive analytics in realizing the energy efficiency and sustainability objectives.

Statistical modeling is also an important aspect of engineering system analysis in addition to energy consumption analysis. Statistical methods give a stringent basis of quantifying uncertainty, modeling variability, and reliability of systems. Aikhuele made a reliance based statistical model to the estimation and optimization of a spur gear system and showed the extent to which statistical formulations may be built into engineering optimization issues successfully [13]. This paper highlights the importance of statistical modeling in the field of engineering, especially in cases where the reliability and interpretation of the results are needed in addition to optimizing the performance. Due to the increasing complexity of engineering problems, nonlinear and high-dimensional optimization problems, the traditional optimization techniques tend to find global optima. In order to overcome this drawback, new optimization algorithms based on mathematical and natural processes have been suggested. Abdel-Basset et al. proposed Exponential Distribution Optimizer which is a metaheuristic algorithm that applies to both global optimization and complex engineering issues [14]. The results of their study show that sophisticated optimization algorithms can be superior to classical ones in the case of non-convex objective functions, and it is clear that evolutionary and population-based algorithms can be used to design system optimization.

The combination of machine learning and statistical modeling has contributed to better predictive engineering functions. In the study of photovoltaic uncertainty in capacitor-planned power systems, Fu suggested a statistical machine learning framework with the aim of explaining how hybrid methods can be useful in accounting uncertainty without significantly reducing the accuracy of prediction [15]. This is especially applicable in energy-related engineering systems where variability and uncertainty are very important in the performance of a system and planning. It also helps to support the importance of mashing up statistical rigor and machine learning flexibility.

The subject of hybrid modeling has been ventured into as well in relation to time-series and the high-frequency engineering data. Alshawarbeh et al. proposed an ARIMA-ANN hybrid architecture of statistical modeling of high-frequency data, which showed better results in comparison with the pure statistical or neural network model [16]. Their research points out the benefits of combining classical statistical models with the artificial intelligence methods to identify both linear patterns and nonlinear ones. These hybrid techniques give a powerful methodology to the intricate engineering systems in which information are mixed in nature.

Besides regression-based and hybrid models, more sophisticated statistical distributions have also been explored to improve the engineering data behavior. Mahmood et al. studied the beta prime distribution and its use in statistical modeling and proved the aptitude of this distribution in the modeling of skewed and heavy tailed data that are often met in engineering and scientific usage [17]. Although this work is not directly concentrated on the energy systems, it adds to the general knowledge of the statistical modeling tools which improve precision and reliability of engineering studies.

Possessing a comparative estimation of the traditional statistic models as compared to the machine learning methods is informative in understanding strengths and weaknesses of various schools of thought of analysis. In a medical engineering setting, Premsagar et al. compared the traditional statistical analysis with machine learning models and found that the latter are more likely to predict well, whereas the former can provide insights [18]. It is a common motif in engineering analytics to trade-off between accuracy and transparency, and is the rational motivation behind the desire to adopt frameworks which combine several modeling approaches instead of applying a single technique.

Topological data analysis is an example of newer methods of analysis with analytical paradigms that have been studied to reveal structural patterns of complex data sets. Madukpe et al. conducted a thorough review of Mapper algorithm and its use in different applications, which is likely to show hidden data structures that cannot be readily identified using more traditional statistical or machine learning tools [19]. Even though these methods are still under developmental stages of application in engineering optimization, the fact that they have emerged is indicative of the continuous development of sophisticated data analytics, as well as the growing number of tools that can be used to analyze engineering systems.

Overall, the selected literature shows that there is a visible tendency towards the incorporation of statistical modeling, machine learning, and advanced optimization methods into the engineering research. Although much has already been done in individual fields, one can notice that the existing researches commonly focus on prediction, modeling, or optimization one by one. It still lacks a methodological advantage in the creation of general frameworks that allow

analytical methods to be systematically compared and incorporated in multi-objective optimization processes. Solutions to this gap are needed in order to further the data-driven engineering practice, especially in energy systems where sustainable and efficient design solutions are becoming more critical.

3. Methodology

3.1 Research Design

The proposed study will use a quantitative and data-driven research design to model and optimize the work of an engineering system based on the most recent methods of data analytics. The methodology is generally a combination of statistical modelling, machine learning and multi-objective optimization in a single analytical model. The research design is explanatory and predictive because it aims at not only understanding the relationship that exists between the system design variables and the performance outcomes but also knowing the best system configurations.

The system of engineering being investigated is residential building energy system, and the performance in this case will be in terms of heating load and cooling load. The research works in four consecutive steps, the first one being data collection and the subsequent data preprocessing process to guarantee consistency and adequacy to prediction, the creation of predictive models using several analytical paradigms, such as statistical and machine learning techniques; the second step is the systematic assessment and comparison of models built with respect to predictive capabilities and generalization; the third step involves the implantation of the models into a multi-objective evolutionary optimization framework to optimise the system design variables and select Pareto-optimal solutions.

Such well-organized methodology makes approaches to be methodologically sound, reproducible and consistent with the aims of smart and sustainable design of engineering systems.

3.2 Data Collection Methods

The research paper provides a combined approach to statistical modeling and optimization of engineering systems with the help of advanced data analytics driven by the increased demand to achieve sustainable and data-driven design methodology when working on energy-intensive applications. The model of energy system in a residential building is used as a typical engineering scenario with the use of Energy Efficiency Dataset [20] on UCI Machine Learning Repository that has 768 simulated building configurations, modeled by eight building geometry and envelope-related variables and two performance measures, heating and cooling loads. Response Surface Methodology, Generalized Additive Models, and Random Forest regression are used together as the methodological framework to model linear, nonlinear, and interaction effect, and systematic model comparison based on a variety of accuracy measures is carried out. The best models are then integrated into a multi-objective optimization problem using the NSGA-II algorithm in order to find Pareto-optimal design solutions.

Findings indicate that machine learning models have predictive accuracy that is better compared to classical statistical methods especially in cooling load, but statistical models have high interpretability and explanatory power. Optimization outcomes show that there is a thin trade-off between cooling and heating demands and hence the energy-efficient design strategies can work to maximize both goals. All in all, actual meaning.

3.3 Population and Sampling

The study population will be all potential residential building designs that take the parameter ranges that the simulation procedure that generated the dataset generates. All observations are associated with a practical design configuration of this population.

The study samples the whole existing data (n = 768) rather than the sample of a larger population, which provides the maximum information and helps to exclude sampling bias. The dataset was divided into training and testing sets with 80:20 split to get 614 observations to be used to train the model and 154 observations to be used to test the model.

Random stratified splitting with a given random seed was used to partition it thus providing reproducibility and consistency across the approaches to modeling. The training and testing divisions were equally allocated to all predictive models to enable fair and unbiased performance comparisons.

This method is consistent with the best practices in research in data-driven engineering and makes reported measures of performance indicators more indicative of actual generalization performance instead of overfitting.

3.4 Data Analysis Techniques

3.4.1 Statistical Modeling: Response Surface Methodology

Response Surface Methodology (RSM) as a base of analysis was used in order to model the correlation between the variables of building design and their energy performance. A 2nd-degree poly-nomial regression equation was developed, using linear terms, quadratic terms and interaction effects. RSM was chosen because of its high interpretability and because its use has been proven in the optimization of engineering research.

Cooling load and heating load were modeled separately. Coefficient of determination (R^2), adjusted R^2, and residual diagnostics were used to determine the model adequacy. The assumptions of linear regression were checked as well as the assumptions that the model represented the behavior of the system in the right statistical way were checked with the help of the residual plots and normal probability plots. Semi-parametric models: Generalized Additive Model Semi-parametric models: Gaussian Discontinuities.

3.4.2 Semi-Parametric Modeling: Generalized Additive Models

Generalized Additive Models (GAMs) were used as a transitional modeling technique to address nonlinear relationships and be able to interpret the model. GAMs change the historical regression models since each predictor can take up the appearance of a smooth function instead of a fixed linear coefficient. Cubic spline smoothers were used in each of the input variables in this study. Penalized likelihood estimation was used to fit GAMs to heating and cooling loads. The test dataset was tested using root mean square error (RMSE), mean absolute error (MAE) and the coefficient of determination (R^2). The GAMs provide a point of contact between classical statistical models and entirely non-parametric machine learning systems, which both can offer information about the nonlinear effects and provide transparency.

3.4.3 Machine Learning Modeling: Random Forest Regression

In case of advanced data analytics, the main machine learning model used was the Random Forest regression. Random Forest is an ensemble technique that tries to build several decision trees and combines their forecasts, which allow strong management of nonlinear connections and intricate interactions among the variables.

The grid search was done with five-fold cross-validation, and the parameters to be optimized were the tree depth, number of estimators, and feature sampling strategy. Hotels Separate random forest models were fitted on heating and cooling loads.

RMSE, MAE, and R^2 were used to measure model performance on the held-out test data. Also, the analysis of the permutation feature importance was performed to measure the relative impact of the individual design variables on the energy performance, and it gave interpretable engineering insights.

3.4.4 Model Comparison and Validation

The three modelling methods were compared systematically with the evaluation metrics being consistent RSM, GAM and Random Forest. The given comparative analysis made it possible to estimate trade-offs between interpretability and predictive accuracy to ensure that the choice of models used to optimize the results was supported by evidence.

The predictive performance of the Random Forest models was demonstrated to be better in both target variables and selection was therefore done as the surrogate models in the next optimization phase.

3.4.5 Multi-Objective Optimization Using NSGA-II

A multi-objective optimization model was applied to determine the best building designs by applying the Non-Dominated Sorting Genetic Algorithm II (NSGA-II). The conflicting objectives of the formulated optimization problem were to minimize the heating load and to minimize the cooling load.

The surrogate objective functions used were the trained Random Forest models, which facilitated the exploration of the design space effectively without necessarily having to simulate the design space computationally. The design variables were constrained in regards to their physical observed ranges whereas the discrete variables were constrained to values that were possible.

NSGA-II was implemented across several generations using a given population size and gave a varied group of Pareto-optimal solutions. The resultant Pareto front was plotted to consequently reveal the representative solutions to minimum heating load, minimum cooling load and balanced performance..

3.5 Methodological Rigor and Reproducibility

All the analyses were done using open-source Python-based libraries and therefore, transparency and reproducibility. The control conditions were the evaluation procedures used in the model, optimization parameters and random seeds. The

combined approach gives it a powerful and generalizable model of statistical modeling and optimization of the engineering systems through the use of the complex data analytics.

4. Results and Analysis

4.1 Data Presentation and Descriptive Overview

The Energy Efficiency Dataset consists of 768 simulated residential building designs, whose design variables are eight and the performance indicators are the heating load (Y1) and cooling load (Y2). Before developing the model, an exploratory analysis was performed to ascertain the presence of no missing values in the dataset and the physical meaning of the ranges of all the variables. The values of heating load are between about 6.0 and 43.1 and the values of cooling load are between about 10.9 and 48.0 meaning there is a great difference in heating and cooling loads depending on the design configurations. This variability offers a robust basis to statistical modelling, machine learning prediction and optimization, as it represents a range of different, and realistic building energy performance conditions.

4.2 Statistical Modeling Results: Response Surface Methodology

4.2.1 Model Performance

Response Surface Methodology (RSM) models were constructed in both cooling load and heating load as the second-order models. Explanatory power was good with RSM especially in heating load. The heating load model was able to produce an R^2 of about 0.995, and the cooling load model was able to produce an R^2 of approximately 0.973. These big values signify that the quadratic associations among building design variables and system performance can be used to clarify a huge portion of the variation in energy demand.

The good results of RSM can be attributed to the deterministic and simulation characteristics of the data set since loads of energy are continuous with regard to geometrical and envelope-related parameters.

4.2.2 Residual Diagnostics

The residual diagnostic tests conducted to assess the adequacy and statistical validity of the Response Surface Methodology (RSM) models were conducted on the basis of residual-versus-fitted plots and normal Q-Q plots of the heating load and cooling load predictions. These Figures show these diagnostics and give a feeling on model behavior, error distribution, and possible deviation of the regression assumption.

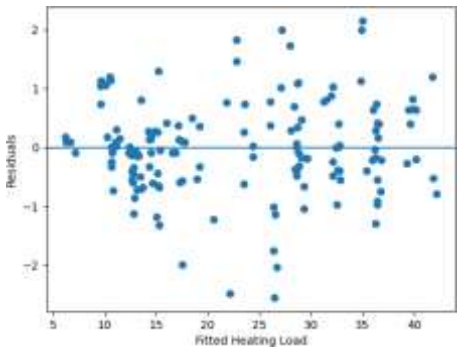

Figure 1. Residuals versus fitted values for the RSM-based heating load prediction model.

The distribution of the residuals is symmetric about zero and does not show any apparent systematic trend implying that the model is capturing the major trends in the data. Although the degree of dispersion is growing as the fitted values go high there is no intense heteroscedasticity.

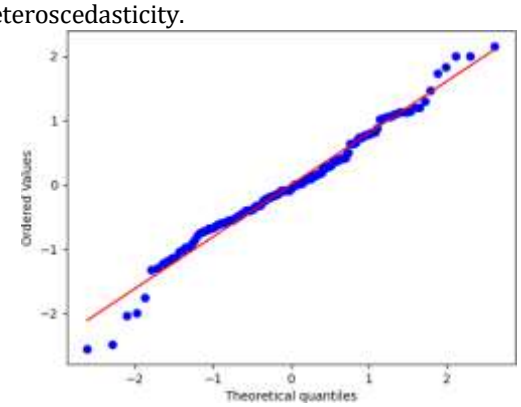

Figure 2. Normal Q–Q plot of residuals for the RSM-based heating load prediction model.

Most of the points are concentrated close to the reference line indicating that indicators of normality are observed in a rough manner with slight deviations at the ends. These deviations are typical of high-accuracy models, and cannot represent important violations of modeling assumptions.

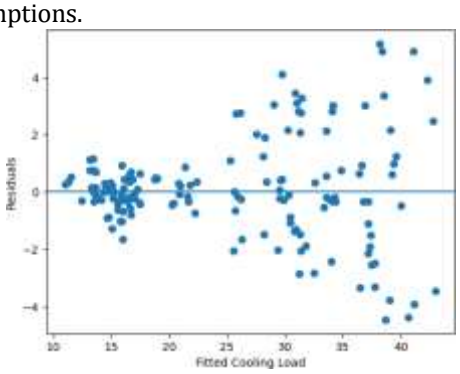

Figure 3. Residuals versus fitted values for the RSM-based cooling load prediction model.

The residual of cooling loads show slightly higher dispersion as compared to heating load especially at high fitted values. This implies that the behavior of cooling loads is a little more complicated and nonlinear, which is in agreement with previous research on the energy systems of buildings.

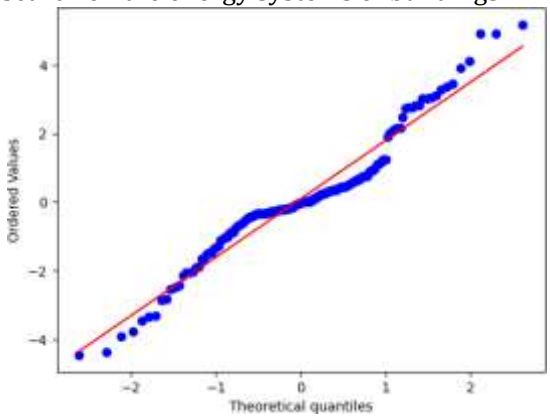

Figure 4. Normal Q–Q plot of residuals for the RSM-based cooling load prediction model.

Whereas the central residuals are also in agreement with the theoretical normal distribution, moderate deviations are noted in the tails. Such deviations rationalize the application of more permissive modeling, e.g. machine learning, in later analysis.In general, the residual diagnostics indicate that RSM offers a statistically valid baseline model and encourages the implementation of sophisticated data analytics tools with an enhanced predictive accuracy.

4.3 Advanced Data Analytics Results

4.3.1 Generalized Additive Model Performance

Generalized Additive Models (GAMs) were used to reduce the linear dependence on the design variables and their relationships with energy performance, and were also used to maintain interpretability. Both targets showed good predictive performance of GAMs, and the pseudo R^2 value of heating load and cooling load was more than 0.98 and 0.96 respectively.

Even though GAMs fare a little worse than RSM with respect to heating load, it offers smooth representation of nonlinear effects and competitive cooling load. Those findings show that even though heating behaviour is dominated by quadratic relationships, nonlinear modeling with flexibility is advantageous to cooling load.

4.3.2 Random Forest Regression Results

The best predictive model between all models was found to be the random Forest regression. Table 1 outlines the predictive performance of RSM, GAM and Random Forest models on the test data in terms of root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2).

Table 1. Comparative predictive performance of modeling approaches

Model	Target	RMSE	MAE	R^2
Random Forest	Heating Load (Y1)	0.49	0.36	0.9977
RSM	Heating Load (Y1)	0.80	0.60	0.9938
GAM	Heating Load (Y1)	1.09	0.81	0.9885
Random Forest	Cooling Load (Y2)	1.60	1.06	0.9724
RSM	Cooling Load (Y2)	1.73	1.19	0.9678
GAM	Cooling Load (Y2)	1.85	1.37	0.9630

Random Forest model had the lowest of predictions errors and highest R^2 values of heating and cooling loads. Cooling load is improved specifically, which supports the fact that ensemble-based machine learning models are more effective to capture complex nonlinear interactions within cooling energy behavior.

4.3.3 Feature Importance Analysis

The feature importance analysis was done using permutation analysis to determine the most important design parameters. Relative compactness, glazing area, roof area and surface area appeared to be the most significant predictors with regard to heating load. Such observations are in line with the principles of the physical heat transfer where building shapes are small and the amount of glazing is minimal to minimize the heat lost.

The cooling load, total building height, roof size and glazing properties were some of the most significant variables to the cooling demand, which indicated the impact of solar exposure, interior air volume, and envelope geometry on cooling demand. The fact that the rankings of model-based importance structures and physical intuition are consistent supports the validity of the machine learning findings.

4.4 Model Comparison and Key Findings

Random Forest had the best overall predictive accuracy, then RSM, and GAM had the best averaged across both targets. Although RSM has shown unexpectedly good results because of the organized character of the data, the Random Forest model has been capable of capturing the remaining nonlinearities that Random Forest is the reason it was selected as the surrogate model to optimize.
A notable methodological observation that the comparative analysis can bring to light is that classical statistical models are still useful in terms of interpretability, although sophisticated machine learning models have a higher predictive accuracy in optimization problems.

4.5 Multi-Objective Optimization Results

4.5.1 Pareto Front Analysis

The trained Random Forest models were used as surrogate functions, and a multi-objective optimization problem was solved with NSGA-II algorithm. These were aimed at reduction of both heating load and cooling load.

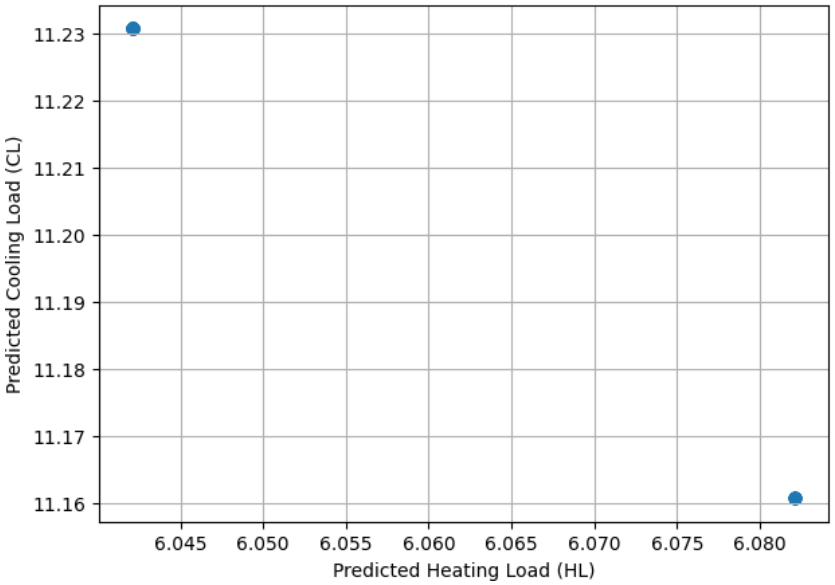

Figure 5. Pareto front obtained using the NSGA-II algorithm for simultaneous minimization of predicted heating load and cooling load in the building energy optimization problem.

The front comprised 120 non-dominated solutions, which means that the optimization process is well converged. It is worth noting that the Pareto front has a slender trade-off area which implies that there is a positive correlation between heating and cooling performance.
This trend suggests that the design configurations that make heating less expensive are also less likely to make cooling expensive, meaning that there is not much conflict among the goals. This is typical of building envelope optimization, where small size, less glazing, and moderate geometry are useful to seasonal energy performance.

4.5.2 Optimal Design Configurations

Three representative solutions were found in the Pareto front, namely minimum cooling load, minimum heating load, and a balanced solution. The minimum heating load design got a predicted heating load of about 6.04 and cooling load of about 11.23. Minimal cooling load design was 11.16 with a slightly high heating load of 6.08.
Interestingly the balanced solution was also the minimum cooling load design thus further confirming the finding that heating and cooling goals are not highly antagonistic. Such ideal designs have similar features such as moderate relative compactness, restrained glazing area, and balanced roof and wall geometry.

4.6 Patterns, Trends, and Implications

There are a number of key patterns in the results. The heating load is always more predictable than cooling load as indicated by a larger coefficient of determination in all the considered models, implying more regular behaviour and structural stability. Machine learning models have better predictive performance over classical statistical methods, and specifically have been shown to be better at cooling load prediction because of their functionality to model complex nonlinear interactions. The energy analysis analysis also shows that energy efficient design strategies are more likely to be beneficial to both the heating and cooling performance of the design such that the Pareto front of the design is narrow, representing that there is no significant conflict between the goals. Lastly, the determined optimal solutions

also can be interpreted physically and aligned with the known engineering principles, which supports the fact that the proposed data-driven optimization framework is practical and applicable.

These results indicate that cutting-edge data analytics can readily assist in sustainable system design of engineering as it allows the correct prediction and informed optimization of complex performance targets.

The findings validate the hypothesis that the suggested integrated model is effective and efficient enough to integrate statistical modeling, advanced data analytics, and multi-objective optimization to examine and enhance the engineering system performance. Besides the high predictive accuracy of the structure, the structure can also be used to give actionable design insights thus its relevance to sustainable engineering practice..

5. Discussion

This study proves the efficiency of the combination of statistical modeling, the high level of data analytics, and multi-objective optimization in the analysis and enhancement of the work of the engineering system. Results of the comparative modeling suggest that there are evident disparities in predictive power among the analytical paradigms which is also indicative of the difference in the capacity of each method to represent nonlinear relationships and interaction of variables that are present in building energy systems. All models provided more accurate predictions of heating load than cooling load indicating that heating demand has more regular and structurally predictable behavior with respect to building design variables. The finding is consistent with the previous results that the heating load in residential buildings is highly controlled by the parameters related to the envelope, including compactness, surface area, and glazing properties, which are generally smoothly varying with the design configurations [12].

Random Forest regression demonstrated the greatest predictive accuracy with regard to heating and cooling loads as indicated by the lowest value of RMSE and greatest value of R^2. It is this high performance that can be explained by the ensemble nature of random forest models that are able to effectively represent nonlinear effects and complex interaction without the need to explicitly assume a functional form. Response Surface Methodology, in turn, even though it showed equally impressive explanatory strength, it proved to have a slightly lower predictive ability to the cooling load. It implies that although quadratic approximations are adequate to explain the main trends in the heating behavior, the cooling demand is affected by higher-order or localized nonlinearities that are better addressed with the help of more flexible modeling methods. Other fields of engineering have also reported similar trade-offs between interpretability and predictive accuracy; machine learning models are more effective in complex nonlinear environments than classical statistical methods [18].

The Generalized Additive Models were in the middle ground in terms of performance with a compromise between flexibility and interpretability. Though GAMs failed to be better predictors than Random Forest models, smooth functional representation offered valuable information about nonlinear effects of variables. This validates the application of semi-parametric models as powerful tools of exploratory analysis and model validation, especially in engineering applications where the information on the impact of the variables is still relevant. The relative performance pecking order in the present investigation aligns with the trends in the hybrid modeling in the energy and power system application, in which machine learning models are able to provide high generalization and yet statistical models are able to provide explanatory value [15].

The analysis of feature importance also supported the physical possibility of the results of modeling. Relative compactness, surface area, roof area and glazing characteristics were found to be predominant predictors of both cooling and heating loads. The results are consistent with the principles of heat transfer and building physics, which affirms that the data-driven models presented in this paper represented significant engineering relationships, and not spurious correlations. The fact that machine learning models can rediscover physically interpretable patterns has been seen as one of the factors that enabled the increasing acceptance of machine learning in the engineering research and practice [9].

The results of the multi-objective optimization give further understanding of the practical implication of the suggested framework. The NSGA-II algorithm was able to find a well-spread Pareto front, which shows that it explored the design space well and non-dominated solutions were reached. It is important to note that the Pareto front had a fairly wide range of trade-offs between heating and cooling goals. This implies that energy

efficient design solutions are more likely to enhance the performance measures at the same time and not create sharp contradictions between the goals. This phenomenon has been noted in previous building energy systems optimization studies in which improvements to the envelope can have year-round payback.

The definition of representative optimal solutions, including minimum heating load, minimum cooling load, and balanced design, proves that the proposed framework has the capability of supporting decisions. The closeness of the balanced solution to the extreme solutions implies that one can easily attain near-optimal performance in both objectives without much trade-off. This is a useful engineering discovery since it eases the decision-making process and makes it possible to implement integrated energy- efficient design approaches. These computational benefits of data-driven optimization methods were made possible by the fact that the optimization of the goal using Random Forest models as surrogate objective functions could easily be performed with no required repeated energy simulations [7].

In comparison with the existing literature, the current study makes a step farther in advancing previous research by providing a systematic and unified comparison of statistical, semi-parametric and machine learning models in one optimization framework. Although earlier works have analyzed the individual modeling methods or paid attention to the accuracy of prediction only, a smaller number have directly considered model appropriateness to downstream optimization. This study fills a major methodological gap of previous studies by introducing predictive models directly into a multi-objective evolutionary algorithm, a crucial methodological shortcoming of existing literature.

Irrespective of its contributions, the study is limited in a number of ways. The simulation-generated data, used to perform the analysis, is not grounded on real-world operation parameters including occupant behavior, adaptive control strategies or stochastic weather variability. Consequently, the results can be viewed as indicating methodological competence and not necessarily giving ideal building designs. Moreover, the optimization goals were confined to heating and cooling loads, but it was not in economic, environmental, or comfort-associated goals that can be potentially important in practice. These restrictions can be aligned with the range of most data-driven engineering research and offer future methodological extensions.

Further studies directions involve the inclusion of more performance goals that can be used to represent the sustainability, e.g., lifecycle cost, carbon emissions, and thermal comfort to provide a more in-depth sustainability assessment. Model realism and relevance would be further increased by the addition of real-world sensor data and dynamically changing environmental inputs. Also, new tools of analysis, such as topological data analysis and deep learning, can provide more information on the complicated set of engineering data and should be considered in the given framework [19].

All in all, the findings validate that the combined application of statistical modeling, high-quality data analytics, and multi-objective optimization are a strong and efficient method of analyzing the engineering systems. The proposed framework will help to achieve sustainable, data-oriented solutions to engineering challenges and ensure the realization of intelligent and sustainable solutions by balancing interpretability, predictive ability and the ability to optimize.

6. Conclusion

The paper introduces a unified model of the statistical modeling and optimization of engineering system with the help of the progressive data analytics which is caused by the increasing necessity of sustainable and data-driven engineering design approaches to energy-intensive systems. A representative example of engineering application is a residential building energy system, heating and cooling loads of which are considered to be functions of the most important geometric and envelope-related design variables, obtained on the basis of a simulated dataset. The methodological framework is a combination of classical Response Surface Methodology, Generalized Additive Models and the Random Forest regression to get the linear, nonlinear and interaction effects, and then a systematic comparison of the predictive performance. The most precise models are then incorporated into a multi-objective optimisation model using the NSGA-II algorithm in order to find Pareto-optimal design configurations that trade off competing energy goals. The findings show that machine learning models are more accurate in predictive performance, especially when it comes to the cooling load, but statistical models have high interpretability. The outcomes of the optimization show that trade-off between heating and cooling requirements is slender, which means that the energy-efficient design solutions can be used to enhance both of them. Generally,

the research demonstrates the importance of statistical analysis and sophisticated data analytics.

References:

[1]. Adeyeye, O., & Akanbi, I. (2024). Optimization in systems engineering: A review of how data analytics and optimization algorithms are applied. *Computer Science & It Research Journal*, *5*(4), 809-823.

[2]. Noman, A. H. M., Mustaquim, S. M., Molla, S., & Siddique, I. M. (2024). Enhancing Operations Quality Improvement through Advanced Data Analytics. *Journal of Computer Science Engineering and Software Testing*, *10*(1), 1-14.

[3]. Sadat Lavasani, M., Raeisi Ardali, N., Sotudeh-Gharebagh, R., Zarghami, R., Abonyi, J., & Mostoufi, N. (2023). Big data analytics opportunities for applications in process engineering. *Reviews in Chemical Engineering*, *39*(3), 479-511.

[4]. Pham, H. (Ed.). (2023). *Springer handbook of engineering statistics*. Springer Nature.

[5]. Oluoha, O. M., Odeshina, A., Reis, O., Okpeke, F., Attipoe, V., & Orieno, O. (2022). Optimizing business decision-making with advanced data analytics techniques. *Iconic Research and Engineering Journals*, *6*(5), 184-203.

[6]. Adesina, A. A., Iyelolu, T. V., & Paul, P. O. (2024). Optimizing business processes with advanced analytics: techniques for efficiency and productivity improvement. *World Journal of Advanced Research and Reviews*, *22*(3), 1917-1926.

[7]. Vagaská, A., Gombár, M., & Straka, Ľ. (2022). Selected mathematical optimization methods for solving problems of engineering practice. *Energies*, *15*(6), 2205.

[8]. Pasupuleti, V., Thuraka, B., Kodete, C. S., & Malisetty, S. (2024). Enhancing supply chain agility and sustainability through machine learning: Optimization techniques for logistics and inventory management. *Logistics*, *8*(3), 73.

[9]. Strielkowski, W., Vlasov, A., Selivanov, K., Muraviev, K., & Shakhnov, V. (2023). Prospects and challenges of the machine learning and data-driven methods for the predictive analysis of power systems: A review. *Energies*, *16*(10), 4025.

[10]. Nwosu, N. T., Babatunde, S. O., & Ijomah, T. (2024). Enhancing customer experience and market penetration through advanced data analytics in the health industry. *World Journal of Advanced Research and Reviews*, *22*(3), 1157-1170.

[11]. Abualigah, L., Zitar, R. A., Almotairi, K. H., Hussein, A. M., Abd Elaziz, M., Nikoo, M. R., & Gandomi, A. H. (2022). Wind, solar, and photovoltaic renewable energy systems with and without energy storage optimization: A survey of advanced machine learning and deep learning techniques. *Energies*, *15*(2), 578.

[12]. Ohalete, N. C., Aderibigbe, A. O., Ani, E. C., Ohenhen, P. E., & Akinoso, A. E. (2023). Data science in energy consumption analysis: A review of AI techniques in identifying patterns and efficiency opportunities. *Engineering Science & Technology Journal*, *4*(6), 357-380.

[13]. Aikhuele, D. (2023). Development of a statistical reliability-based model for the estimation and optimization of a spur gear system. *Journal of Computational and Cognitive Engineering*, *2*(2), 168-174.

[14]. Abdel-Basset, M., El-Shahat, D., Jameel, M., & Abouhawwash, M. (2023). Exponential distribution optimizer (EDO): a novel math-inspired algorithm for global optimization and engineering problems. *Artificial Intelligence Review*, *56*(9), 9329-9400.

[15]. Fu, X. (2022). Statistical machine learning model for capacitor planning considering uncertainties in photovoltaic power. *Protection and Control of Modern Power Systems*, *7*(1), 1-13.

[16]. Alshawarbeh, E., Abdulrahman, A. T., & Hussam, E. (2023). Statistical modeling of high frequency datasets using the ARIMA-ANN hybrid. *Mathematics*, *11*(22), 4594.

[17]. Mahmood, R. S., Mizban, R. J., Sarhan, M. A., Rashid, A., RASHEED, M., & Saidani, T. Analysis and applications of the beta prime distribution in statistical modeling. *Journal of Positive Sciences*, *3*(6), 34-41.

[18]. Premsagar, P., Aldous, C., Esterhuizen, T. M., Gomes, B. J., Gaskell, J. W., & Tabb, D. L. (2022). Comparing conventional statistical models and machine learning in a small cohort of South African cardiac patients. *Informatics in Medicine Unlocked*, *34*, 101103.

[19]. Madukpe, V. N., Ugoala, B. C., & Zulkepli, N. F. S. (2026). A comprehensive review of the mapper algorithm, a topological data analysis technique, and Its applications across various fields (2007–2025). *International Journal of Data Science and Analytics*, *21*(1), 56.

[20]. A. Tsanas and A. Xifara. "Energy Efficiency," UCI Machine Learning Repository, 2012. [Online]. Available: https://doi.org/10.24432/C51307.