Data driven Urban Building Energy Modeling with Machine Learning in Satom CH

Ahad Montazeri
Department of Energy, DENERG
Politecnico di Torino
Torino, Italy
ahad.montazeri69@outlook.com

Jérôme H. Kämpf
Energy Informatics Group
Idiap Research Institute
Martigny, Switzerland
jerome.kaempf@idiap.ch

Guglielmina Mutani
Department of Energy, DENERG
Politecnico di Torino
Torino, Italy
guglielmina.mutani@polito.it

***Abstract*—This article delves into the integration of district heating systems into urban planning for sustainable development in regions with moderate to cold climates. The study introduces the Data-driven Urban Energy modeling framework, which aims to bridge the gap between conventional engineering-based energy simulation models and emerging data-driven machine learning (ML) models. By doing so, it provides accurate and comprehensive insights into urban energy demand (ED) patterns.**

The methodology involves evaluating engineering and ML model's generalization power, revealing its ability to predict energy demand accurately at both building and urban scales. Machine learning algorithms, including LightGBM (LGBM) and Random Forest (RF) regression, are employed to fine-tune the energy-use model for future energy demand predictions. The results demonstrate the model's exceptional accuracy and suitability for diverse urban scenarios. Incorporating a more straightforward approach like Multiple Linear Regression (MLR) into the methodology also highlights its capability to predict energy demand in less complex research scenarios and offer valuable insights for effective urban energy planning.

Overall, this article emphasizes the significance of data-driven approaches and machine learning techniques in optimizing energy demand, promoting sustainable urban development, and guiding informed decision-making for energy-efficient cities. The findings have implications for urban planners, policymakers, and energy analysts seeking to enhance energy efficiency and contribute to a greener and more sustainable future for urban communities.

Keywords— Urban building energy modeling, Data-driven models, Machine learning, Place-based approach, Geographic Information System GIS

I. INTRODUCTION

In numerous countries characterized by moderate to cold climates, the substantial portion of overall energy consumption stems from the need for space heating and domestic hot water in buildings [1]. While district heating and cooling systems have been in existence for some time, they are continually evolving worldwide [2]. Presently, integrating these systems into urban planning has become a crucial component of transitioning cities towards low-carbon and sustainable development. On one hand, in densely populated residential and commercial areas, district heating and cooling systems demonstrate superior energy efficiency compared to individual systems, promoting energy-saving endeavors [3]. On the other hand, these systems offer an opportunity to leverage local resources, fostering economic development within the region [4]. The ability to make precise and efficient energy demand predictions is essential for achieving the objectives of evaluating new building design alternatives and optimizing energy systems [5].

Advancements in sensing technologies and the rise of smart city initiatives have resulted in an abundance of structured and unstructured data streams describing buildings and the urban environment. In parallel, artificial intelligence is rapidly developing new machine learning models that utilize this data to predict and characterize various physical phenomena within cities [6]. The primary aim of this research is to introduce an innovative and cutting-edge methodology called the Data-driven Urban Energy Modeling framework. This framework seeks to address the disparity between conventional engineering-based energy simulation models and the rapidly evolving data-driven machine learning models [6].

In contrast to physical models, data-driven models eliminate the need for building thermal balance equations, resulting in reduced or no reliance on detailed building physical information. These data-driven models rely on historical data to uncover the underlying connections between output variables, such as building energy demand, and input variables, such as weather conditions, building characteristics, occupant behaviors, and equipment schedules, through the application of mathematical methods [7].

The Data-driven Urban Energy modeling framework incorporates state-of-the-art machine learning techniques to analyze and interpret the complex data streams. By doing so, it can capture intricate patterns, relationships, and dependencies that traditional models might overlook. This data-driven approach not only enhances the accuracy of energy predictions but also provides valuable insights into the factors influencing energy demand patterns in urban areas. Furthermore, the framework is designed to be adaptable and scalable, ensuring its applicability to diverse urban settings and energy-related challenges.

II. CASE STUDY

Satom SA, a pioneering energy company, has implemented an innovative and environmentally friendly solution that harnesses the energy contained in biomass waste and plastics. The company utilizes this abundant waste material to produce water vapor, which, in turn, drives a steam turbine, generating clean and renewable electricity. This process not only mitigates the harmful environmental impacts of waste disposal but also contributes to a sustainable energy future.

An integral part of Satom's energy system is its waste heat recovery mechanism. Traditionally, a significant amount of heat is lost to the environment through thermodynamic

processes involved in thermal machines. However, Satom has successfully developed a thermal network that captures steam withdrawn from the steam turbine and heat collected from the fumes to redistributes this waste heat. By utilizing the thermal network, the recovered heat is efficiently channeled into surrounding buildings, providing them with an eco-friendly heating solution. This approach not only maximizes energy efficiency of the thermodynamic cycle but also significantly reduces the overall carbon footprint of the community.

The distribution of thermal energy is facilitated by an extensive underground network, strategically designed to cover a wide perimeter in the municipalities of Collombey-Muraz and Monthey. This remarkable thermal infrastructure enables Satom to extend its heating services to multiple dwellings, spanning several kilometers, and cater to the energy needs of a substantial population.

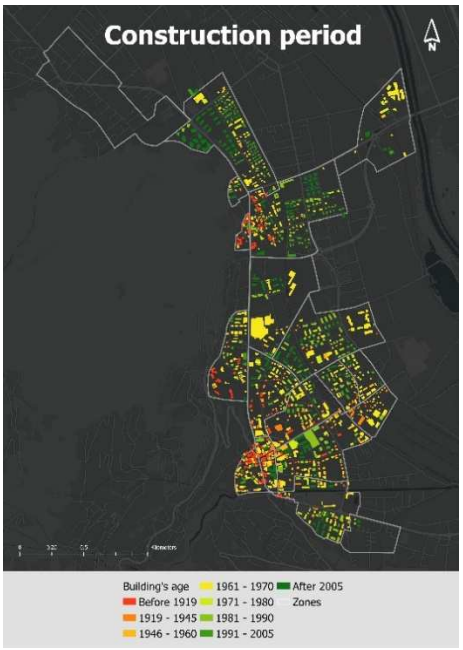

Fig. 1. Construction period of buildings within DHN zone

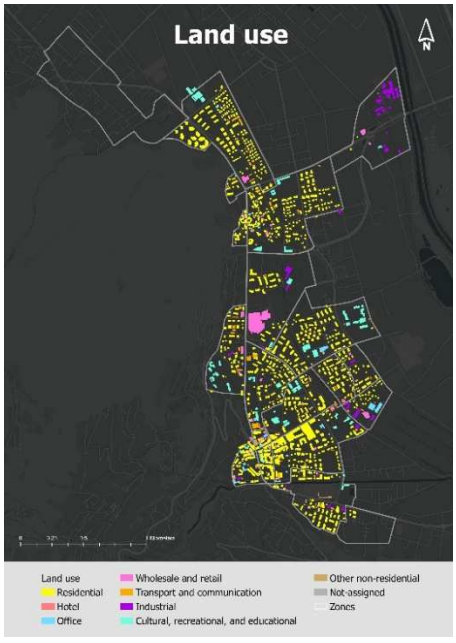

Fig. 2. Land use of buildings within DHN zone

Technically, Satom's energy system characterized by 91 GWh of annual energy injection into the network. The network constitutes of 79 kilometers of pipes including 484 powered substations. This power and size of network allows Satom SA to heat 850,000 m^2 of living space. The research incorporated two primary constraints, namely technical and economical constraints. The technical aspect focused on buildings within the zone where the District Heating Network (DHN) actively provides heating services. To address the economical constraint, residential buildings with S/V values below 0.8 m^{-1} and non-residential buildings below 1.0 m^{-1} were included. The zone comprised a total of 1286 buildings (approximately 5.93 Mm^3), out of which 995 (approximately 4.94 Mm^3) were deemed economically feasible for connection to the network. Fig. 1 depicts the age distribution of buildings, and Fig. 2 presents the various types of buildings based on their usage. Within this valid cluster of buildings, there were 855 residential buildings (approximately 4.06 Mm^3) and 140 non-residential buildings (approximately 0.88 Mm^3), with 355 (approximately 2.68 Mm^3) and 58 (approximately 0.5 Mm^3) buildings successfully connected, respectively. The combined annual heating demand for all connected buildings of both types amounts to approximately 41.57 GWh.

It is important to highlight that the volume distribution of residential buildings underwent statistical testing, resulting in the division of the buildings into two categories. The first category consists of 816 valid buildings with total volume of approximately 3.17 Mm^3. The second category includes 39 so-called abnormal residential buildings with a volume of 0.89 Mm^3. These abnormal buildings will be treated differently in the predictions of energy demand.

III. Methodology

The core methodology utilized in this research centers on Urban Building Energy Modeling, which involves employing place-based techniques to simulate and assess the energy performance of buildings situated in urban environments with a district heating network. By considering various factors such as building age, construction materials, occupancy patterns, weather conditions, and energy usage types, this approach allows for a comprehensive understanding of the energy consumption patterns and potential efficiency improvements within the urban building stock [8, 9].

The research employed advanced data analysis methods to enhance energy efficiency and predictive capabilities within urban environments (Fig. 3). To achieve this, the study first established a comprehensive analysis of the existing engineering-based energy simulation models used in urban planning and energy management. Subsequently, the research delved into the emerging field of data-driven machine learning models, exploring their potential applications and benefits for urban energy simulations [10].

The physical model was constructed in ArcGIS, utilizing surveyor's data and multipatch files from swissBUILDINGS3D versions 2.0 and 3.0 beta. This integrated approach allowed for a comprehensive analysis of the heating demands and potential connections in the studied area. To determine the heating demand of connected buildings, the research utilized the XML model of the case study in the CitySim Pro platform, incorporating the climate file from the year 2021.

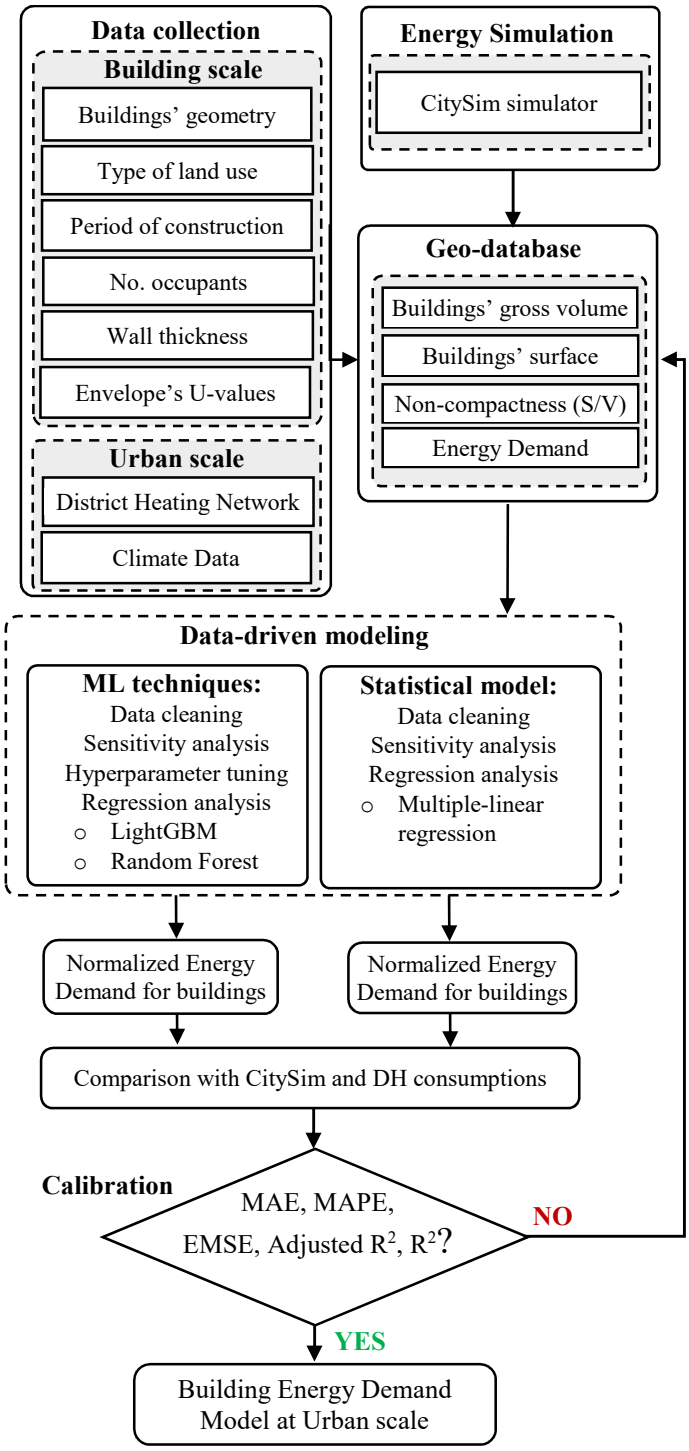

Fig. 3. Flow chart of the methodology: Data-driven Urban Building Energy Modeling

Following the clustering of buildings based on their typology and construction period, a normal distribution analysis is applied to identify and exclude outliers, ensuring the formation of a robust and reliable database. The refined database is then leveraged to implement multilinear regression techniques, which play a pivotal role in constructing an energy-use model specifically tailored to the characteristics of the buildings.

To comprehensively evaluate and compare the performance of machine learning models in energy-use modeling, two prominent algorithms, LightGBM (LGBM) and Random Forest (RF) regression, were employed. By using both models, this study aimed to assess their suitability and effectiveness in capturing the complex relationships between building attributes and energy demand patterns.

LGBM represents a highly efficient implementation of the gradient boosting algorithm, a machine learning technique that employs an ensemble of weak learners, usually decision trees, to address regression or classification tasks. Unlike other ensemble algorithms, gradient boosting adds weak learners to the model sequentially, with each learner being fitted to the residuals of the previous one. This sequential addition of learners enhances the model's predictive power and overall performance [11].

RF is a supervised learning algorithm used for regression tasks, employing a bagging (bootstrap aggregating) approach. It builds upon decision trees, creating multiple trees to obtain average prediction results. In the prediction process of Random Forest, each decision tree is randomly formed with different features and training datasets, allowing them to be trained in parallel. This parallelization contributes to higher prediction accuracy compared to a single decision tree. Moreover, Random Forest effectively addresses overfitting concerns by establishing multiple decision trees, each working on a random sample of the original dataset [12, 13, 14].

LGBM and RF were chosen due to their proven capabilities and widespread use in various domains, including regression tasks. Through rigorous experimentation and analysis, the performance and predictive accuracy of each model were thoroughly assessed. This allowed for a detailed understanding of how well the models captured the underlying patterns in energy demand, and which model exhibited better generalization to unseen data. For hyperparameter tuning of the machine learning models, the Optuna optimizer was employed to tune the LGBM model, while genetic algorithms were utilized to fine-tune the RF model.

Optuna is an open-source optimization software that utilizes a define-by-run API, enabling users to dynamically construct the parameter search space. It offers efficient implementation of searching and pruning strategies, and its versatile architecture allows for various applications, from scalable distributed computing to lightweight interactive experiments. In the context of hyperparameter optimization, Optuna frames the process as minimizing or maximizing an objective function that takes a set of hyperparameters as input and returns a validation score. Unlike traditional methods using static variables, Optuna dynamically constructs the search space within the objective function itself. Each optimization process is called a "study", and each evaluation of the objective function is referred to as a "trial". As the objective function interacts with the trial object, it progressively builds the search space during runtime. Inside the objective function, the user employs the "suggest API" to dynamically generate hyperparameters for each trial. When "suggest API" is invoked, a hyperparameter is statistically sampled based on the historical data from previously evaluated trials [15].

Genetic Algorithm (GA) is an optimization technique inspired by natural selection principles. It operates as a population-based search algorithm, incorporating the concept of "survival of the fittest". By applying genetic operators iteratively on individuals within the population, new populations are generated. The crucial components of GA include chromosome representation, selection, crossover,

mutation, and fitness function computation [16]. In the context of hyperparameter tuning, genetic algorithms explore different combinations of hyperparameters, crossbreeding and mutating them to generate new and potentially better solutions. Table I illustrates the main input data with their source and tools used to process them.

TABLE I. MAIN INPUT DATA IN BUILDING OF ENERGY-USE MODELS

Input data	*Source*	*Tools*
Construction period	The surveyor's data	GIS
Use type	The surveyor's data	GIS
Building's volume	swissBUILDINGS3D_2.0	GIS
Building's surface	calculated	GIS
S/V ratio	swissBUILDINGS3D_2.0	GIS
Building's occupants	XML file for SATOM	None
Wall thickness	XML file for SATOM	None
Net heated surface	calculated	GIS
U_{roof}, U_{wall}, U_{slab}	XML file for SATOM	None
Annual Heating Demand	XML file for SATOM – Climate file 2021	CitySim Pro

IV. DISCUSSION AND RESULTS

The analysis commenced by developing an energy-use model specifically tailored for residential buildings. This model was constructed using linear regression and took into account the age of buildings and their compactness, represented by the S/V (Surface-to-Volume) ratio. Based on the distribution of the buildings' S/V ratios, two distinct typologies were identified: buildings with S/V ratios of 0.5 or lower and those with S/V ratios exceeding 0.5.

The analysis was carried out on two separate clusters of buildings. However, for the cluster with low-density buildings (S/V > 0.5), the frequency was relatively limited after conducting a normal distribution check and removing any outliers. Consequently, the energy-use model was unable to accurately predict energy demand for all buildings in this particular cluster.

To address this limitation and ensure a comprehensive energy-use model, MLR was employed. This approach incorporated all connected residential buildings, taking into consideration all available parameters. The model was then fine-tuned, ensuring that only statistically significant parameters were included in the regression. By doing so, the energy-use model was enhanced to provide a more comprehensive and accurate representation of energy demand for all connected residential buildings.

The MLR model's performance (Fig. 4) was thoroughly assessed to measure the generalization ability of the developed energy-use model when applied to unseen data. The analysis involved plotting the results on graphs to provide a visual representation of the model's predictive capabilities.

The findings indicated that the MLR model demonstrated moderate performance in predicting normalized ED at the building scale. While it exhibited relatively accurate predictions on an individual building level, the model showcased even better accuracy when predicting ED not only at the building scale but also at the larger urban scale.

According to the results of the ML R model, the ED for the entire set of valid residential buildings in the zone amounts to 51.10 GWh per year. Out of these buildings, 324 buildings (with a total volume of 1.96 Mm^3) are already connected and contribute to the mentioned ED. Additionally, there are 492 buildings (with a total volume of 1.21 Mm^3) that have the potential to be connected and would contribute to an ED of 21.21 GWh per year.

Subsequently, the analysis progressed to incorporate machine learning algorithms, specifically LightGBM and RF, to derive the energy-use model and evaluate their accuracy and suitability for future ED predictions.

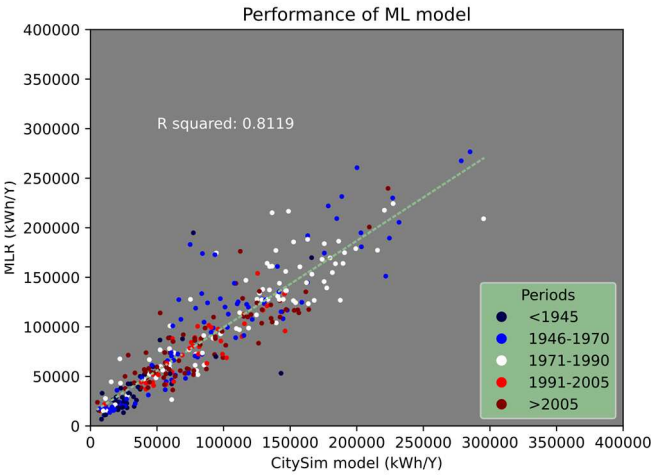

Fig. 4. ED prediction power of MLR at building scale

The generalization capabilities of these ML models specifically for the valid residential buildings were examined. Both LightGBM and RF were utilized aiming to capture the complex relationships between various building attributes and ED patterns.

The assessment of the LightGBM and RF models (R^2: 0.98 and 0.91 respectively) demonstrates that the developed energy-use models exhibit exceptionally high generalization power when applied to unseen data. This remarkable generalization extends to predictions of normalized ED, both at the building scale and the urban scale. Furthermore, the models display exceptional accuracy in forecasting annual ED across various scenarios where RMSE for LightGBM and RF is 1.59 and 3.25 respectively (Fig. 5 & Fig. 6).

However, despite the impressive generalization power exhibited by LightGBM, it is worth noting that the RF model seemed to be a bit weaker in this particular research setting. RF is known to perform exceptionally well with large-size datasets, but in this study, the sample database was limited in size. As a result, the performance of the RF model was somewhat affected, showing relatively lower accuracy compared to LightGBM. Nevertheless, the developed energy-use model still demonstrated remarkable accuracy in forecasting annual ED across various scenarios (refer to Fig.8).

According to the outcomes from the LightGBM model, the ED for all valid residential buildings within the zone is 48.09 GWh per year, and the potentially connectable buildings would require an additional 18.39 GWh per year. However, the RF model suggests slightly different values, with the ED for all residential buildings amounting to 48.63 GWh per year (+1.1%), and for the connectable buildings, it is 18.81 GWh per year (+2.3%). Table II presents the computed ED for the entire stock of buildings, including both those currently connected and those that can be connected in the future, using various regression models that were tested.

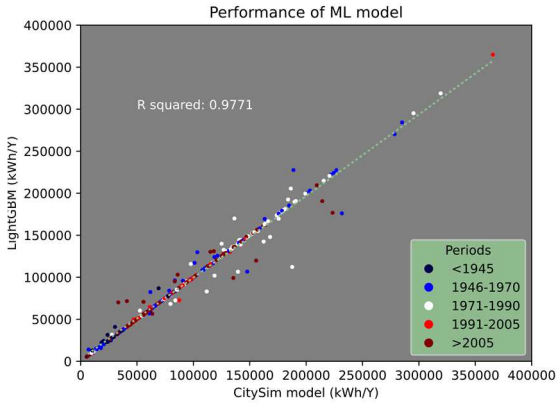

Fig. 5. Energy Demand prediction power of LightGBM at building scale

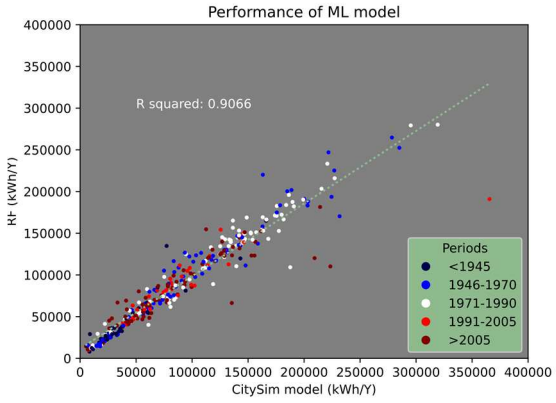

Fig. 6. Energy Demand prediction power of Random Forest Regression at building scale

TABLE II. ESTIMATING ENERGY DEMAND USING DIVERSE METHODS FOR ALL USAGE TYPES

Type of buildings		Model	Whole buildings	Connected buildings	Connectable buildings
Valid residential	No.	--	816	324	492
	Volume (Mm^3)	--	3.17	1.96	1.21
	Share of volume (%)	--	100	61.83	38.17
	ED (GWh/Y)	CitySim Pro	--	29.84	--
		MLR	51.1	29.89 (+0.2%)	21.21
		LightGBM	48.09	29.7 (-0.5%)	18.39
		RF	48.63	29.82 (-0.1%)	18.81
Abnormal residential	No.	--	39	31	8
	Volume (Mm^3)	---	0.89	0.72	0.17
	Share of volume (%)	-	100	80.90	19.1
	ED (GWh/Y)	CitySim Pro	--	8.5	--
		LightGBM	10.74	8.66 (+1.9%)	2.08
		RF	11.56	9.31 (+9.5%)	2.25
Non-residential	No.	--	140	58	82
	Volume (Mm^3)	--	0.88	0.5	0.38
	Share of volume (%)	--	100	55.82	43.18
	ED (GWh/Y)	CitySim Pro	--	5.93	--
		LightGBM	10.02	5.73 (-3.4%)	4.29
		RF	10.56	5.92 (-0.2%)	4.64

The findings suggest (Fig. 7 & Fig. 8) that the MLR model shows promise in accurately predicting ED for individual buildings and urban areas as a whole. Its capacity to provide precise ED forecasts at various scales offers reliable benefits for urban planners and energy analysts. On the other hand, the energy-use model, utilizing LightGBM and RF algorithms, effectively captures underlying data patterns, enabling it to make highly precise predictions even for unseen data. The model's high generalization power ensures reliability in diverse scenarios, making it a valuable tool for accurately forecasting ED for buildings and entire urban settings.

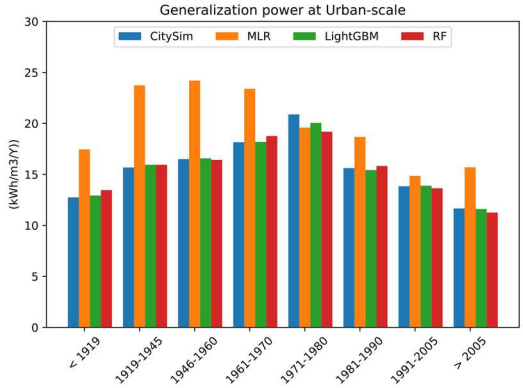

Fig. 7. Normalized Energy Demand prediction power of MLR, LightGBM and RF at urban scale

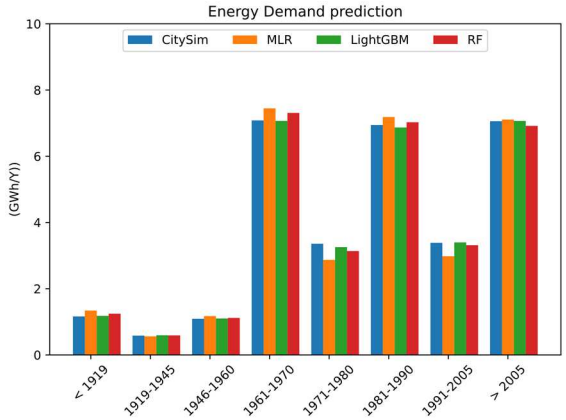

Fig. 8. Energy Demand prediction that uses MLR, LightGBM and RF models at urban scale

V. CONCLUSION

In conclusion, this research successfully employed advanced data analysis methods to enhance energy efficiency and predictive capabilities within urban environments. Through the integration of engineering-based energy simulation models and data-driven machine learning techniques, the study developed a comprehensive energy-use model specifically tailored for buildings and urban areas.

The multiple linear regression model showed promise in accurately predicting ED for individual buildings and urban areas as a whole. Its capacity to provide precise ED forecasts at various scales and its ease of use is an essential benefit for urban planners and energy analysts in energy management.

Furthermore, to thoroughly evaluate the performance of machine learning models, LightGBM and Random Forest Regression were chosen and subjected to rigorous experimentation and analysis. The results showed that both models exhibited high generalization power, accurately predicting ED at the building scale and larger urban scale. Additionally, the energy-use model utilizing these machine learning algorithms displayed exceptional accuracy in forecasting annual ED across various scenarios.

In closing, this research contributes significantly to the field of urban energy efficiency and predictive modeling. By combining engineering-based simulation models with data-driven machine learning techniques, it offers a powerful tool for accurately forecasting ED in buildings and urban settings. The findings and methodologies presented in this study can inform and guide future urban planning and energy management efforts, ultimately leading to more sustainable and energy-efficient cities. Given the paper's highly effective methodology in forecasting urban ED, it will be applied in future research to conduct a comprehensive analysis aimed at evaluating diverse energy-saving scenarios. This analysis seeks to optimize the performance of district heating networks while also exploring the potential for predicting far-future DHN expansion, capitalizing on the advantages derived from energy-saving strategies.

ACKNOWLEDGMENT

We would like to extend our heartfelt gratitude to Satom SA for their invaluable support and generous assistance in our research endeavors. Their willingness to share vital information and insights played a pivotal role in the successful execution of our study. Their dedication to environmental stewardship and sustainable energy practices is commendable, and we are deeply appreciative of the information provided by Satom SA which greatly enriched our research and significantly contributed to the depth and accuracy of our findings. Their collaboration underscores the importance of industry-academic partnerships in advancing knowledge and fostering innovative solutions.

REFERENCES

[1] Sperling, K., & Möller, B. (2012). End-use energy savings and district heating expansion in a local renewable energy system–A short-term perspective. Applied energy, 92, 831-842.

[2] Rezaie, B., & Rosen, M. A. (2012). District heating and cooling: Review of technology and potential enhancements. Applied energy, 93, 2-10.

[3] Lake, A., Rezaie, B., & Beyerlein, S. (2017). Review of district heating and cooling systems for a sustainable future. Renewable and Sustainable Energy Reviews, 67, 417-425.

[4] Dou, Y., Sun, L., Fujii, M., Kikuchi, Y., Kanematsu, Y., & Ren, J. (2021). Towards a renewable-energy-driven district heating system: key technology, system design and integrated planning. Renewable-Energy-Driven Future, 311-332.

[5] Chen, Y., Chen, Z., Xu, P., Li, W., Sha, H., Yang, Z., ... & Hu, C. (2019). Quantification of electricity flexibility in demand response: Office building case study. Energy, 188, 116054.

[6] Nutkiewicz, A., Yang, Z., & Jain, R. K. (2017). Data-driven Urban Energy Simulation (DUE-S): Integrating machine learning into an urban building energy simulation workflow. Energy Procedia, 142, 2114-2119.

[7] Chen Y, Guo M, Chen Z, Chen Z, Ji Y. (2022). Physical energy and data-driven models in building energy prediction: A review. Energy Reports. Nov 1;8:2656-71.

[8] Guelpa, E., Mutani, G., Todeschi, V., Verda V. (2017). A feasibility study on the potential expansion of the district heating network of Turin, Energy Procedia 122, CISBAT 2017, pp. 847-852, 10.1016/j.egypro.2017.07.446.

[9] Mutani, G., Todeschi, V., Guelpa, E., & Verda, V. (2020). Building Efficiency Models and the Optimization of the District Heating Network for Low-Carbon Transition Cities. Springer Proceedings in Energy. Cham. 10.1007/978-3-030-31459-0_14.

[10] Mutani, G., Vocale, P., & Javanroodi, K. (2023). Toward Improved Urban Building Energy Modeling Using a Place-Based Approach. Energies 16, 3944, pp.1-17, 10.3390/en16093944.

[11] Todeschi, V., Boghetti, R., Kämpf, J. H., & Mutani, G. (2021). Evaluation of Urban-scale building energy-use models and tools—Application for the city of Fribourg, Switzerland. Sustainability, 13(4), 1595, 10.3390/su13041595.

[12] Dudek, G. (2015). Short-term load forecasting using random forests. In Intelligent Systems' 2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014, September 24-26, 2014, Warsaw, Poland, Volume 2: Tools, Architectures, Systems, Applications (pp. 821-828). Springer International Publishing.

[13] Lahouar, A., & Slama, J. B. H. (2015). Day-ahead load forecast using random forest and expert input selection. Energy Conversion and Management, 103, 1040-1051.

[14] Moon, J., Kim, Y., Son, M., & Hwang, E. (2018). Hybrid short-term load forecasting scheme using random forest and multilayer perceptron. Energies, 11(12), 3283.

[15] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

[16] Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. Multimedia tools and applications, 80, 8091-8126.