Integrating Building Survey Data with Geospatial Data: A Cluster-Based Ethical Approach

Vidisha Chowdhury[1], Gabriela Gongora-Svartzman[2*], Erin Trochim[3], Philippe Schicker[2]
[1]The Wharton School of the University of Pennsylvania, Philadelphia, PA, USA
[2]Heinz College, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Alaska Center for Energy and Power, University of Alaska Fairbanks, Fairbanks, AK, USA
[1]vidishac@wharton.upenn.edu, [2*]ggongora@cmu.edu, [3]edtrochim@alaska.edu

Abstract

This research paper delves into the unique energy challenges faced by Alaska, arising from its remote geographical location, severe climatic conditions, and heavy reliance on fossil fuels while emphasizing the shortage of comprehensive building energy data. The study introduces an ethical framework that leverages machine learning and geospatial techniques to enable the large-scale integration of data, facilitating the mapping of energy consumption data at the individual building level. Utilizing the Alaska Retrofit Information System (ARIS) and the USA Structures datasets, this framework not only identifies and acknowledges limitations inherent in existing datasets but also establishes a robust ethical foundation for data integration. This framework innovation sets a noteworthy precedent for the responsible utilization of data in the domain of climate justice research, ultimately informing the development of sustainable energy policies through an enhanced understanding of building data and advancing ongoing research agendas. Future research directions involve the incorporation of recently released datasets, which provide precise building location data, thereby further validating the proposed ethical framework and advancing efforts in addressing Alaska's intricate energy challenges.

1 Introduction

Alaska faces unique energy challenges stemming from its remote location, extreme weather, and reliance on fossil fuels [1, 2]. One key barrier is the lack of granular, large-scale data on building energy demands, which is essential to decarbonization efforts in supporting climate change mitigation. While previous efforts have manually collected heating load measurements, these studies cover only small areas. In contrast, the Alaska Retrofit Information System (ARIS) [3, 4] contains a much more extensive database of home energy audits. ARIS was recently released for public use, however, records have been de-identified in terms of geolocation to only zip codes to protect personal and private information.

This paper introduces a novel framework combining machine learning and geospatial techniques to map ARIS home energy data at the building level, at scale. The framework presents an ethical approach for integrating the ARIS database with housing attributes from the USA Structures dataset [5]. Through clustering and validation, the framework accurately matches records across datasets while protecting privacy, using clustering to match records privately. The contributions of this paper are threefold: 1) highlighting current limitations in key energy data sources; 2) providing an ethical framework for integrating such datasets at scale using machine learning; and 3) demonstrating how this approach can inform sustainable energy policy through improved understanding of building energy loads.

Ethical considerations are paramount when utilizing geospatial data on real individuals for climate justice [6]. Researchers must minimize privacy risks and socio-environmental disparities. Proper consent, anonymization, communication, and assessing long-term impacts are critical for ethical data use aligned with equity. This project grappled with ethically combining home energy and

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

housing data lacking clear common IDs. Overall, this work exemplifies how artificial intelligence and geospatial big data can further renewable energy goals while prioritizing transparency and fairness.

2 Previous Work

Traditional geospatial analysis methods have played a role in understanding spatial patterns and making informed decisions in various domains, including energy resource management; however, these methods have limitations [7]**.** In contrast, cloud-based geospatial analytics techniques, such as Google Earth Engine (GEE), have emerged as a leading technique in geospatial analysis [8]–[12], offering major data archives and enabling users to perform complex geospatial tasks at scale [13]–[15] including analysis of renewable energy potential and energy infrastructure changes [16]–[22]. Unsupervised machine learning techniques (ML), such as K-means clustering, have gained importance when uncovering complex geospatial insights regarding energy analytics challenges [23]–[25]. K-means clustering has been applied to energy research [26]–[28] to identify consumption patterns and group regions with similar energy demand profiles [29]–[31]. On the other hand, supervised ML, such as ensemble learning (e.g., Random Forest and XGBoost), have shown promise in predicting energy consumption and optimizing energy efficiency in various applications [18], [32]–[37].

Despite the advances in machine learning techniques [38] , current research addressing heating or cooling loads at a residential level only focuses on improving machine learning prediction power [39]–[41], implementing IoT devices to measure city resources [42], and claim to be data-driven, but fail to address transparency and ethical concerns regarding the existing datasets. Furthermore, current research aims at collecting more data without considering already existing historical and governmental datasets and without considering the people behind the datasets. This paper proposes a framework to join datasets from the Alaska Retrofit Information System (ARIS) database on energy-retrofit residences, along with data extracted from GEE, by ethically considering each variable as an affected individual and its future ramifications. The authors of this paper hope to bring insights into future research addressing the problem with existing Alaska energy datasets and, therefore, contribute to the field of energy analytics and resource management.

3 Framework

The proposed framework (Figure 1) was developed with the primary objective of joining administrative survey data to geospatial building data in an ethical manner, wherein exact building locations are not used for the joins. Table 1 (Appendix A) provides a description of the ARIS Housing datasets and the cleaning they underwent, including the removal of identical duplicates and addressing out-of-range values, with ‘YearBuilt’ missing values imputed using iterative regression and K-Nearest Neighbor (KNN) approaches. Separately, Table 2 (Appendix A) describes the USA Structures dataset. In steps 1, 2 and 3 of Figure 1,various geospatial datasets were combined to estimate the base area, height and age of each building footprint in the USA Structures data. The process of height estimation gave rise to negative height values for around 8% of buildings which further had to be replaced by imputed values from the distribution of heights for the remaining buildings. In step 4, zip codes and borough information was added to the USA Structures dataset. This was followed by KMeans clustering based on building features (area, height and age) within each Alaskan borough (step 5). The Calinski-Harabasz index and Silhouette scores helped determine the optimal number of clusters and each building in a borough was assigned a cluster ID. In step 6, the cluster ID was treated as an output variable and a decision tree was fitted using building area, height and age as features, for each borough. Subsequently in step 7, the ARIS data set of home energy audits was cleaned and the same set of building features (area, height and age) along with energy consumption and zip codes were obtained from it. In step 8, clusters were predicted for each building in ARIS using the model estimated for its borough in step 6. For ARIS buildings, a building’s borough was obtained from its zip code.

Steps 9, 10 and 11 describe how every building in ARIS was matched to the most similar building in its zip code from USA Structures based on Euclidean distances (cosine similarity and Manhattan distance were among the other distance metrics computed) between vectors of building features (area, height and age) for every pair of buildings. This process was repeated for every building in each borough.

In the absence of exact building location data, the clusters and zip codes served as filters to reduce the search space for a building's match and make the matching process more efficient. The clusters represent how similar buildings are in their characteristics, conditional on belonging to the same zip code. Clustering was done at the borough level because a borough is a large enough geographical boundary that has a sufficient number of buildings of various types. Filtering buildings by borough was equivalent to conditioning on approximate location which made our matches more accurate without sacrificing too much statistical power. Further, buildings within boroughs roughly experience similar climates which will become more relevant when we predict energy needs of buildings based on changes in climate in the future.

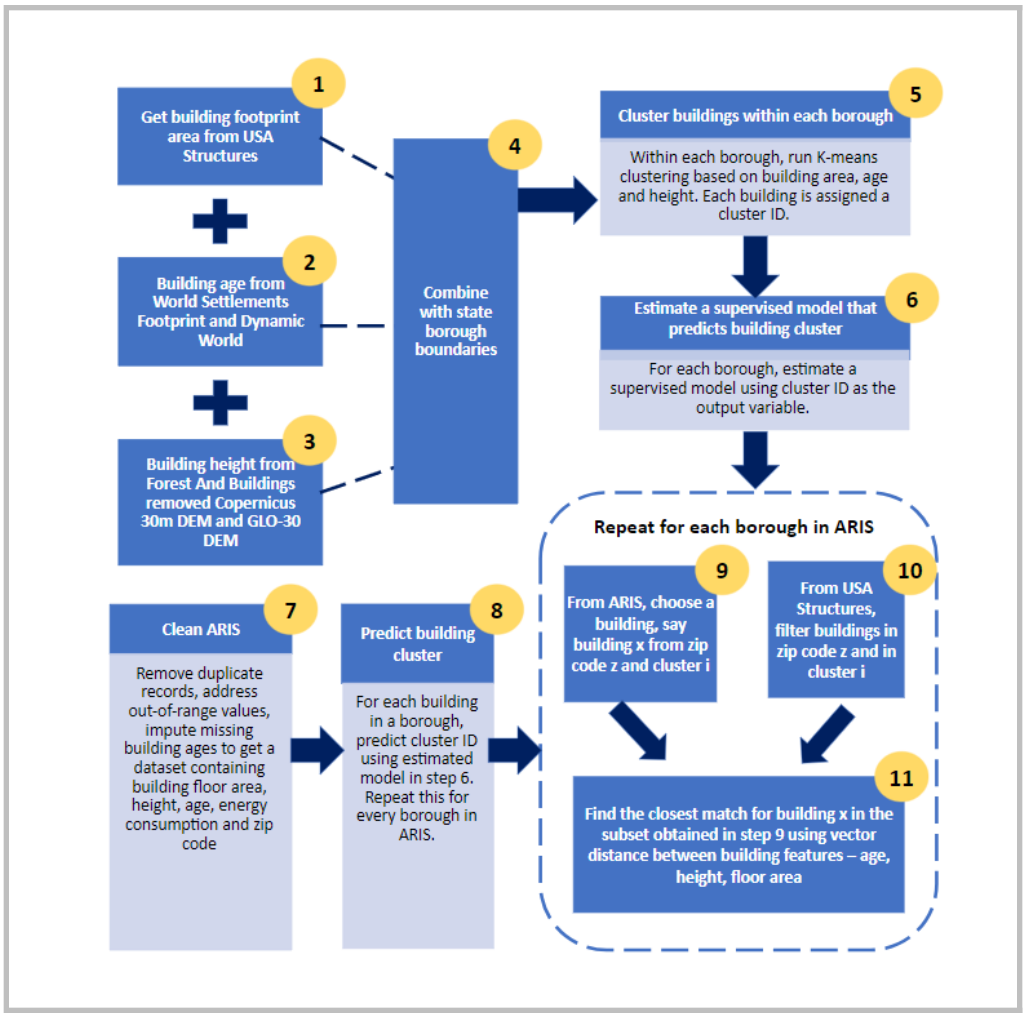

Figure 1: Overall Framework describing the initial geospatial data (1-3), analysis at borough boundary (3), cluster analysis of geospatial data (5-6), ARIS data (7) and clustering (8), and dataset matching (9-11).

4 Results and Validation

The framework explained in the previous section was applied to each borough in Alaska. In the results presented here, the matching of buildings is based on the Euclidean distance metric between vectors of building characteristics (area, height, age) from the two data sets. Validating our matching approach can prove to be a challenge due to limited data on building characteristics for the whole of Alaska. Further, given that most of such data are de-identified in terms of geolocation, we can only use aggregate distributions of building features to validate our matching framework. One approach is to compare the distributions of building features from ARIS to the corresponding distributions from geospatial data in the final matched dataset. The matched pairs of features from each dataset can be plotted in the same graph.

The specific borough presented as a case study from the proposed framework is the Denali borough. The Denali borough has a total of 590 buildings spread over twelve thousand square miles. Figure 2 shows the results of three important features from the joins made on the Denali Borough. Figure 2)b) shows the closest match in the distribution of area (in square feet) for all the buildings in both ARIS and USA Structure datasets. This indicates that the framework proposed in this paper has closely matched buildings in terms of base area. The height in Figure 2)c) has the same shape between the USA Structures and ARIS density plots. This again indicates that the framework is working and suggests measurement errors in the collection of the datasets.

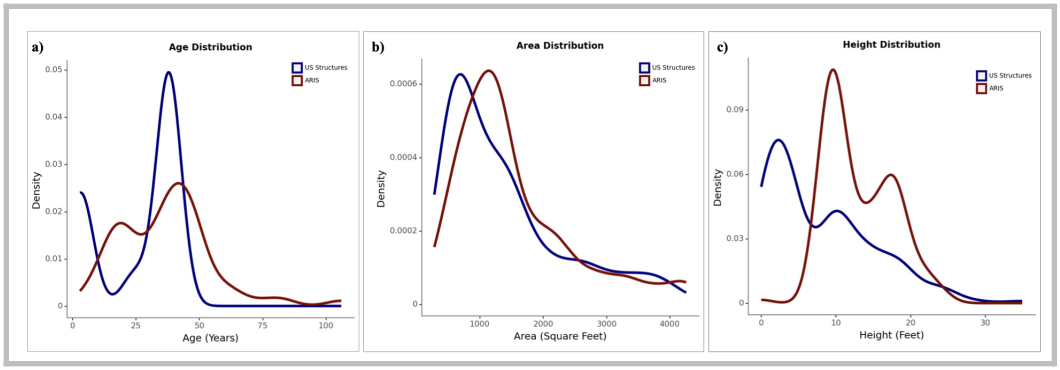

Figure 2: Denali Borough, AK showing age (a), area (b) and height (c) distributions.

The age of buildings is an important feature when considering future heating loads models, and unfortunately, it was a feature poorly recorded in both datasets. Figure 2)a) shows the age of the buildings having dissimilar densities. The ARIS recorded heights show a multimodal density plot, which would indicate that this data was recorded for different groups of buildings or at different times. This makes sense, given the nature of the ARIS dataset. For the USA structures, the density plot shows a high peak, indicating a high number of buildings built around the mean age reported. For the ARIS dataset, 3% of the observations in the 'YearBuilt' variable had missing values, which were imputed using a KNN approach; this justifies the variability of its density plot. For the USA structure dataset, missing building ages were assigned a specific pre-agreed value based on the nature of the geospatial data set being used to obtain ages.

In the future, we intend to incorporate the BlocPower [43] and Model America [44] datasets for validating our building matches and energy demand estimates. These datasets have been recently released for research purposes and include precise building location data. However, through this process, we also intend to highlight the tradeoffs between accuracy in estimation and using sensitive features like location in estimation. The ultimate goal behind our ethical framework is to show how close we can get to true energy demand values without utilizing sensitive information like exact locations and compromising on privacy.

Our final dataset consists of buildings across the state of Alaska with their corresponding annual and hourly energy demands. This can be used to inform current policies aimed at improving the energy efficiency of buildings in the state. For instance, our data can quantify the tradeoff between retrofitting households and replacing old construction by giving approximate energy costs in both scenarios and facilitate an economically efficient decision.

5 Conclusions and Discussions

This paper delves into the distinctive energy challenges faced by Alaska due to its remote geographical positioning and severe climate conditions, all while emphasizing the inadequacy of comprehensive building energy data. Within this context, it introduces a rigorous ethical framework designed to facilitate the large-scale integration of data - with the help of machine learning and geospatial techniques - for the purpose of mapping energy consumption data at the individual building level, drawing from both the Alaska Retrofit Information System (ARIS) and the USA Structures dataset. The contributions of this research encompass the identification and acknowledgment of limitations inherent to existing datasets, as well as the formulation and establishment of a robust ethical data integration framework. This approach sets a precedent for the responsible utilization of data within the realm of climate justice research, ultimately serving to inform the development of sustainable energy policies through enhanced comprehension of building data and the progression of ongoing research considerations. Future endeavors will expand upon this framework by incorporating the BlocPower [43] and Model America [44] datasets, recently released for research purposes, which include precise building location data. These datasets will help us validate our proposed ethical framework and our energy demand estimates. Most importantly, such an expansion would advance ethical considerations pertaining to sensitive features, thereby representing a significant stride toward addressing Alaska's intricate energy challenges.

6 Acknowledgements

We would like to express our gratitude to Maddie Gaumer, Shamsi Soltani, and Nicholas Bolten for their help during the Summer of 2022, as part of the Data Science for Social Good Fellowship at the University of Washington.

This work was supported by the Broad Agency Announcement Program from the U.S. Army Cold Regions Research and Engineering Laboratory (ERDC-CRREL) under contract No. W913E521C0017 from the U.S Army Basic Research Program (Program Element 0603119A, Ground Advanced Technology).

References

[1] Alaska Energy Authority, ‘Renewable Energy Fund (REF) Grants’. Alaska Energy Authority. [Online]. Available: https://www.akenergyauthority.org/What-We-Do/Grants-Loans/Renewable-Energy-Fund

[2] Municipality of Anchorage, ‘Anchorage Climate Action Plan’, Anchorage, AK, May 2019. [Online]. Available: https://www.muni.org/Departments/Mayor/AWARE/ResilientAnchorage/pages/climateactionplan.aspx

[3] Alaska Housing Finance Corporation, ‘Housing Energy Efficiency AHFC Energy Programs and Resources’, [Online]. Available: https://www.energy.gov/sites/prod/files/2016/02/f30/16_bob_brean_tues0429.pdf

[4] Alaska Housing Finance Corporation, ‘Empowering Alaska Borrowers With Information To Save Money’, Dec. 2022. [Online]. Available: https://www.ahfc.us/blog/posts/empowering-alaska-borrowers-information-save-money

[5] Esri_US_Federal_Data, ‘USA Structures’. Federal Emergency Management Agency (FEMA) Geospatial Response Office, Oak Ridge National Laboratory (ORNL). [Online]. Available: https://services2.arcgis.com/FiaPA4ga0iQKduv3/arcgis/rest/services/USA_Structures_View/FeatureServer

[6] P. Schicker, S. Soltani, M. Gaumer, V. Chowdhury, N. Bolten, and E. Trochim, ‘Environmental Justice Considerations for Remote Sensing Approaches: Calculating Heating Loads in Alaska through Geospatial and Machine Learning Techniques’, vol. 2022, pp. SY44C-01, Dec. 2022, Accessed: Sep. 30, 2023. [Online].

[7] J. Zhang and M. F. Goodchild, *Uncertainty in geographical information*. London: Taylor & Francis, 2003.

[8] M. Gaumer, N. Bolten, V. Chowdhury, P. Schicker, S. Soltani, and E. Trochim, ‘Estimating Heating Loads in Alaska using Remote Sensing and Machine Learning Methods’.

[9] V. Chowdhury, S. Soltani, P. Schicker, M. Gaumer, N. Bolten, and E. Trochim, ‘Informing Decarbonization through Machine Learning (ML): A Geospatial ML Pipeline for Estimating Heating Loads’, vol. 2022, pp. GC14A-02, Dec. 2022, Accessed: Sep. 30, 2023. [Online].

[10] Q. Zhao, L. Yu, X. Li, D. Peng, Y. Zhang, and P. Gong, ‘Progress and Trends in the Application of Google Earth and Google Earth Engine’, *Remote Sens.*, vol. 13, no. 18, p. 3778, Sep. 2021, doi: 10.3390/rs13183778.

[11] A. V. Uzhinskiy, ‘Google Earth Engine and machine learning for Earth monitoring’, in *Proceedings of The 6th International Workshop on Deep Learning in Computational Physics — PoS(DLCP2022)*, JINR, Dubna, Russia: Sissa Medialab, Nov. 2022, p. 021. doi: 10.22323/1.429.0021.

[12] M. I. Habibie, ‘The Applications of Machine Learning using Google Earth Engine for Remote Sensing Analysis’, *J. Teknoinfo*, vol. 16, no. 2, p. 233, Jul. 2022, doi: 10.33365/jti.v16i2.1872.

[13] O. Mutanga and L. Kumar, ‘Google Earth Engine Applications’, *Remote Sens.*, vol. 11, no. 5, p. 591, Mar. 2019, doi: 10.3390/rs11050591.

[14] M. Amani *et al.*, ‘Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review’, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5326–5350, 2020, doi: 10.1109/JSTARS.2020.3021052.

[15] Avtar *et al.*, ‘Exploring Renewable Energy Resources Using Remote Sensing and GIS—A Review’, *Resources*, vol. 8, no. 3, p. 149, Aug. 2019, doi: 10.3390/resources8030149.

[16] S. Soltani, V. Chowdhury, P. Schicker, M. Gaumer, N. Bolten, and E. Trochim, ‘A Novel Geospatial-First Machine Learning Approach to Modeling Heating Loads in Alaska: When Perfect (Data) are the Enemy of Good (Data)’, vol. 2022, pp. GH23B-08, Dec. 2022, Accessed: Sep. 30, 2023. [Online].

[17] H. Binfei, H. Xujun, H. Mingguo, L. Yitao, and L. Shiwei, ‘Research Progress on the

Application of Google Earth Engine in Geoscience and Environmental Sciences', *Remote Sens. Technol. Appl.*, vol. 33, no. 4, pp. 600–611, 2018, [Online]. Available: http://www.rsta.ac.cn/EN/10.11873/j.issn.1004-0323.2018.4.0600

[18] Y. Li, C. Liu, J. Zhang, P. Zhang, and Y. Xue, 'Monitoring Spatial and Temporal Patterns of Rubber Plantation Dynamics Using Time-Series Landsat Images and Google Earth Engine', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 9450–9461, 2021, doi: 10.1109/JSTARS.2021.3110763.

[19] H. H. Jaafar, R. M. Mourad, W. P. Kustas, and M. C. Anderson, 'A Global Implementation of Single‐ and Dual‐Source Surface Energy Balance Models for Estimating Actual Evapotranspiration at 30‐m Resolution Using Google Earth Engine', *Water Resour. Res.*, vol. 58, no. 11, p. e2022WR032800, Nov. 2022, doi: 10.1029/2022WR032800.

[20] B. A. Wong, C. Thomas, and P. Halpin, 'Automating offshore infrastructure extractions using synthetic aperture radar & Google Earth Engine', *Remote Sens. Environ.*, vol. 233, p. 111412, Nov. 2019, doi: 10.1016/j.rse.2019.111412.

[21] R. Martins Moreira, A. Conceição Paranhos Filho, and S. Sieber, 'A novel Water-Food-Energy nexus approach integrating Analytic Hierarchy Process and Google Earth Engine using global datasets for photovoltaic energy generation', *Renew. Energy Focus*, vol. 43, pp. 210–227, Dec. 2022, doi: 10.1016/j.ref.2022.09.001.

[22] H. Supe *et al.*, 'Google Earth Engine for the Detection of Soiling on Photovoltaic Solar Panels in Arid Environments', *Remote Sens.*, vol. 12, no. 9, p. 1466, May 2020, doi: 10.3390/rs12091466.

[23] I. K. Nti, J. A. Quarcoo, J. Aning, and G. K. Fosu, 'A mini-review of machine learning in big data analytics: Applications, challenges, and prospects', *Big Data Min. Anal.*, vol. 5, no. 2, pp. 81–97, Jun. 2022, doi: 10.26599/BDMA.2021.9020028.

[24] H. Hamdoun, A. Sagheer, and H. Youness, 'Energy time series forecasting-analytical and empirical assessment of conventional and machine learning models', *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, pp. 12477–12502, Jun. 2021, doi: 10.3233/JIFS-201717.

[25] S. Ardabili, L. Abdolalizadeh, C. Mako, B. Torok, and A. Mosavi, 'Systematic Review of Deep Learning and Machine Learning for Building Energy', *Front. Energy Res.*, vol. 10, p. 786027, Mar. 2022, doi: 10.3389/fenrg.2022.786027.

[26] J. A. Hartigan and M. A. Wong, 'Algorithm AS 136: A K-Means Clustering Algorithm', *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.

[27] A. Likas, N. Vlassis, and J. J. Verbeek, 'The global k-means clustering algorithm', *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, Feb. 2003, doi: 10.1016/S0031-3203(02)00060-2.

[28] K. P. Sinaga and M.-S. Yang, 'Unsupervised K-Means Clustering Algorithm', *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[29] J. Sachs, D. Moya, S. Giarola, and A. Hawkes, 'Clustered spatially and temporally resolved global heat and cooling energy demand in the residential sector', *Appl. Energy*, vol. 250, pp. 48–62, Sep. 2019, doi: 10.1016/j.apenergy.2019.05.011.

[30] J. Bejarano *et al.*, 'Sampling Within k-Means Algorithm to Cluster Large Datasets', ORNL/TM-2011/394, 1025410, Aug. 2011. doi: 10.2172/1025410.

[31] B. Dash, D. Mishra, A. Rath, and M. Acharya, 'A hybridized K-means clustering approach for high dimensional dataset', *Int. J. Eng. Sci. Technol.*, vol. 2, no. 2, pp. 59–66, Sep. 2010, doi: 10.4314/ijest.v2i2.59139.

[32] D. Fawzy, S. Moussa, and N. Badr, 'The Evolution of Data Mining Techniques to Big Data Analytics: An Extensive Study with Application to Renewable Energy Data Analytics', *Asian J. Appl. Sci.*, vol. 4, no. 3, Jun. 2016, [Online].

[33] F. Wahid, R. Ghazali, A. S. Shah, and M. Fayaz, 'Prediction of Energy Consumption in the Buildings Using Multi-Layer Perceptron and Random Forest', *Int. J. Adv. Sci. Technol.*, vol. 101, pp. 13–22, Apr. 2017, doi: 10.14257/ijast.2017.101.02.

[34] Y.-T. Chen, E. Piedad, and C.-C. Kuo, 'Energy Consumption Load Forecasting Using a Level-Based Random Forest Classifier', *Symmetry*, vol. 11, no. 8, p. 956, Jul. 2019, doi: 10.3390/sym11080956.

[35] P. C. Sen, M. Hajra, and M. Ghosh, 'Supervised Classification Algorithms in Machine Learning: A Survey and Review', in *Emerging Technology in Modelling and Graphics*, vol.

937, J. K. Mandal and D. Bhattacharya, Eds., in Advances in Intelligent Systems and Computing, vol. 937. , Singapore: Springer Singapore, 2020, pp. 99–111. doi: 10.1007/978-981-13-7403-6_11.
[36] J. Dhanalakshmi, N. Ayyanathan, and N. S. Pandian, 'Energy Analytics and Comparative Performance Analysis Of Machine Learning Classifiers On Power Boiler Dataset', in *2019 11th International Conference on Advanced Computing (ICoAC)*, Chennai, India: IEEE, Dec. 2019, pp. 72–78. doi: 10.1109/ICoAC48765.2019.246819.
[37] M. W. Ahmad, M. Mourshed, and Y. Rezgui, 'Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption', *Energy Build.*, vol. 147, pp. 77–89, Jul. 2017, doi: 10.1016/j.enbuild.2017.04.038.
[38] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, and I. G. Maglogiannis, 'Supervised machine learning: A review of classification techniques.', in *Emerging Artificial Intelligence Applications in Computer Engineering*, IOS Press, 2007, pp. 3–24. [Online].
[39] Z. Wei *et al.*, 'Prediction of residential district heating load based on machine learning: A case study', *Energy*, vol. 231, p. 120950, Sep. 2021, doi: 10.1016/j.energy.2021.120950.
[40] X. Li and R. Yao, 'A machine-learning-based approach to predict residential annual space heating and cooling loads considering occupant behaviour', *Energy*, vol. 212, p. 118676, Dec. 2020, doi: 10.1016/j.energy.2020.118676.
[41] Y. Hossain, P. A. Loring, and T. Marsik, 'Defining energy security in the rural North—Historical and contemporary perspectives from Alaska', *Energy Res. Soc. Sci.*, vol. 16, pp. 89–97, Jun. 2016, doi: 10.1016/j.erss.2016.03.014.
[42] R. Chaganti *et al.*, 'Building Heating and Cooling Load Prediction Using Ensemble Machine Learning Model', *Sensors*, vol. 22, no. 19, p. 7692, Oct. 2022, doi: 10.3390/s22197692.
[43] Environmental Impact Data Collaborative, 'BlocPower Active [Alpha+]'. Sep. 21, 2023. [Online]. Available: https://redivis.com/datasets/c8kf-fwz3md6rs?v=1.1
[44] J. New, M. Adams, A. Berres, B. Bass, and N. Clinton, 'Model America – data and models of every U.S. building'. Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States). Oak Ridge Leadership Computing Facility (OLCF); Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States);Argonne National Laboratory (ANL) Leadership Computing Facility (ALCF), 2021. doi: 10.13139/ORNLNCCS/1774134.
[45] M. Marconcini *et al.*, 'World Settlement Footprint (WSF) 2015'. figshare, p. 2688996507 Bytes, 2020. doi: 10.6084/M9.FIGSHARE.10048412.V1.
[46] M. Marconcini, A. Metz- Marconcini, T. Esch, and N. Gorelick, 'Understanding Current Trends in Global Urbanisation - The World Settlement Footprint Suite', *GI_Forum*, vol. 1, pp. 33–38, 2021, doi: 10.1553/giscience2021_01_s33.
[47] C. F. Brown *et al.*, 'Dynamic World, Near real-time global 10 m land use land cover mapping', *Sci. Data*, vol. 9, no. 1, p. 251, Jun. 2022, doi: 10.1038/s41597-022-01307-4.
[48] L. Hawker *et al.*, 'A 30 m global map of elevation with forests and buildings removed', *Environ. Res. Lett.*, vol. 17, no. 2, p. 024016, Feb. 2022, doi: 10.1088/1748-9326/ac4d4f.

Appendix A - Data Description

Table 1 shows the datasets within the ARIS Housing data used. There were several considerations taken when cleaning these datasets, which are essential to note. Due to the possibility of the same building being surveyed more than once and in order to retain building characteristics for each building, identical duplicates were dropped from the dataset. Several variables in ARIS had values outside the plausible range (e.g., the 'YearBuilt' had values greater than the current year); therefore, these values were removed and treated as missing values. YearBuilt is an important feature for future models but had 2.5% missing values; thereby, the missing values were imputed using an iterative regression approach as well as a K-Nearest Neighbor (KNN) approach.

Table 1: ARIS Housing Data

Name	Description
ARIS Fuel Data	Dataset describing fuel consumption for each building with features including different fuel types (e.g., coal, gas). Each building has a unique ProjectId.
ARIS Building Ratings	Dataset describing building features such as floor area, ceiling height, number of bedrooms, and more. Each building has a unique ProjectId.

Table 2 shows the datasets and processes taken to calculate the building's height, area, and age. Once these calculations were done separately, the data was merged into one complete dataset. Features with more than 90% missing values were dropped. Approximately 8% of buildings in the data had negative height values due to imprecision in the GEE height calculation. Their heights were imputed (and replaced) by height values following the distribution of height values from the rest of the data.

Table 2: Geospatial Data (USA Structures)

Task	Datasets	Process
Building age	-World Settlements Footprint Data (1985-2015) (WSF) [45] -Updated WSF layer (2019) (WSF2019) [46] -Dynamic World dataset (June 2015- present) [47]	The median building age was calculated for each footprint in the WSF dataset. Buildings with no associated age were checked against the WSF2019 and Dynamic World datasets. The age of a building was then either updated or imputed. If imputation was not possible, the building age was assigned to 1985. A single dataset with building age estimates was created for each footprint in the state of Alaska.
Building height and area	- FABDEM (Forest And Buildings removed Copernicus 30m DEM) [48] -GLO-30 (Copernicus 30m Digital Elevation Model) -Zipcode data	Iterate over each zip code, calculating the height of buildings by reducing the building height image over the filtered building collection in each zip code to compute the average building height. Calculate the area of each building and add it as a feature to the dataset. Combine the distinct zip codes into one feature collection and export this singular dataset.