Estimating Building Energy Efficiency From Street View Imagery, Aerial Imagery, and Land Surface Temperature Data

Kevin Mayer
Department of Civil and Environmental Engineering
Stanford Unversity
kdmayer@stanford.edu

Lukas Haas
Department of Computer Science
Stanford University
lukashaas@stanford.edu

Abstract

In the race towards carbon neutrality, the building sector has fallen behind and bears the potential to endanger the progress made across other industries. This is because buildings exhibit a life span of several decades which creates substantial inertia in the face of climate change. This inertia is further exacerbated by the scale of the existing building stock. With several billion operational buildings around the globe, working towards a carbon-neutral building sector requires solutions which enable stakeholders to accurately identify and retrofit subpar buildings at scale. However, improving the energy efficiency of the existing building stock through retrofits in a targeted and efficient way remains challenging. This is because, as of today, the energy efficiency of buildings is generally determined by on-site visits of certified energy auditors which makes the process slow, costly, and geographically incomplete. In order to accelerate the identification of promising retrofit targets, this work proposes a new method which can estimate a building's energy efficiency using purely remotely sensed data such as street view and aerial imagery, OSM-derived footprint areas, and satellite-borne land surface temperature (LST) measurements. We find that in the binary setting of distinguishing efficient from inefficient buildings, our end-to-end deep learning model achieves a macro F1-score of 62.06%. While this means that the end-to-end deep learning model outperforms the k-NN and SVM-based baseline models by 5.62 to 11.47 percentage points respectively, the ablation study reveals that averaging the predictions of models trained on different data sources boosts model performance by more than 1.6 percentage points to an F1-score of 63.67%. As such, this work shows the potential and complementary nature of remotely sensed data in predicting building attributes such as energy efficiency and opens up new opportunities for future work to integrate additional data sources.

1. Introduction

Accounting for the whole building life cycle, from construction to usage, renovation, and demolition, the EU estimates that buildings are responsible for 40% of the union's energy consumption and for 36% of its greenhouse gas emissions. Yet, as of today, about 75% of the EU building stock remains energy inefficient and, on average, less than 1% of the national building stocks are renovated each year. This means that in order to meet its own energy and climate objectives, the EU needs to at least double its current rate of renovations over the next years [4].

However, the renovation rate in itself is not a good indicator for the decarbonization of the built environment. This is because in order to maximize the efficiency gains from building retrofit programs as quickly and cost-optimized as possible, the buildings with the largest retrofit potential, in general those with the poorest energy performance, need to be retrofitted first. Although more than 80 states around the world have already developed building energy codes and many require building energy performance certificates whenever a building is sold or rented, large-scale and publicly available datasets on building-level energy performance are scarce. This scarcity in information creates a bottleneck in the transformation of the built environment as it remains difficult for actors in the retrofit industry to identify promising retrofit targets automatically and across geographies.

To identify buildings with high retrofit potential on a large scale, this work uses an end-to-end deep learning model which fuses different input data sources such as street view images, aerial images, and land surface temperature data in order to predict building-level energy performance from remotely sensed data only. We evaluate our approach by conducting a real-world case study which spans around 40,000 buildings across four diverse cities in the United Kingdom. The model's predictions are evaluated using entries from the UK's official building energy performance registry.

2. Related Work

Using machine learning-based methods to estimate building characteristics such as energy consumption [20, 23, 8, 22] and efficiency [13, 24], photovoltaic rooftop potential [15, 14] and generation [21, 19], as well as property type, age, and value [12, 1] has received significant research attention. In general, these studies can be further sub-divided into **top-down** approaches which start with estimates for a whole city or region and disaggregate them as needed and **bottom-up** approaches which in turn focus on individual buildings first [7]. Since this paper estimates the energy efficiency for individual buildings, the subsequent review focuses on bottom-up approaches.

2.1. Bottom-up Approaches for Energy Consumption and Efficiency

Predicting the energy consumption and efficiency of buildings is important because it can inform utility companies, residents, facility managers, contractors, and public agencies on how to improve the energy efficiency of the building stock.

With the emergence of the first city-scale building-level benchmark datasets, earlier studies have focused on using tabular data to predict building energy consumption and efficiency. Incorporating information such as the building area, age, and the number of floors, [13] presents a regression-based approach for commercial buildings with more than 50,000 square feet. Similarly, [27] develops a random forest model to predict energy-related building characteristics using tabular features such as a building's surface, wall, and roof area in order to estimate its heating load (HL) and cooling load (CL).

In contrast to the studies relying on tabular data, [20, 2] utilize historical consumption data collected with smart meters in order to predict a building's short-term energy usage. While [20] relies on a random forest-based model to estimate hourly building energy consumption, [2] focuses on the comparison and evaluation of different modeling techniques, ranging from artificial neural networks, over support vector machines, to linear regression, and tree-based methods.

Due to the lack of large-scale and publicly available building energy datasets, a growing body of research is studying building energy consumption and efficiency from remotely sensed data. Unlike previous approaches which are inherently limited to small geographic regions, the increasing availability of high-resolution remotely sensed data empowers this stream of research to potentially scale across geographies. In [23], the authors make use of overhead aerial imagery with a spatial resolution of 0.3m and publicly available building footprint information in order to derive estimates for residential energy consumption in a three-step procedure. After detecting and segmenting buildings in the overhead imagery, the authors classify the buildings by type into commercial and residential properties. Lastly, the building energy consumption is predicted with a random forest-based model which takes image-derived building features, i.e. footprint area and perimeter as well as the building type, as input. On a building-level, the model achieves an R^2 of 0.28 and 0.38 for the case studies in Gainesville and San Diego, respectively. [8] extends this line of work in two ways. First, their work collects and analyzes two overhead images at different zoom-levels per building in order to better understand a building's spatial context. Second, the authors also generate a building-specific context vector based on the establishments within a given radius R. This fixed length context vector intends to capture the potential type of occupancy based on the social function of nearby establishments. In an end-to-end deep learning model, three distinct neural network branches are used to independently project the context vector and the two images into a joint feature space. After combining the embedded feature vectors in the joint feature space, another five-layer feed-forward neural network maps the embedded features to a real-valued energy consumption estimate per building. Similarly, [22] models building electricity consumption solely based on aerial and street view images. By adding street view images, the authors are able to achieve results which are comparable to conventional models based on public tabular datasets. As in [23], the authors find that spatially aggregating the predictions further improves the results.

Apart from the studies that focus on building energy consumption, [24] presents a model which uses street view imagery and tabular data such as a building's total floor area, height, and number of open fireplaces in order to estimate a building's energy efficiency on a scale from A-G, a rating scheme introduced according to the EU's directive on the energy performance of buildings (EPBD), with "A" being the most energy efficient and "G" being the least. In a real-world case study for the city of Glasgow, more than 30,000 buildings are analyzed and the model achieves an accuracy of 86.8%.

While previous studies have used a combination of aerial and street view images to estimate building energy consumption, methods to estimate building energy efficiency from purely remotely sensed data have yet to be developed. Moreover, previous studies also do not include satellite-borne heat loss information derived from long wave infrared measurements. Hence, we extend the existing literature by shifting the focus to a new combination of purely remotely sensed data sources in an end-to-end deep learning model.

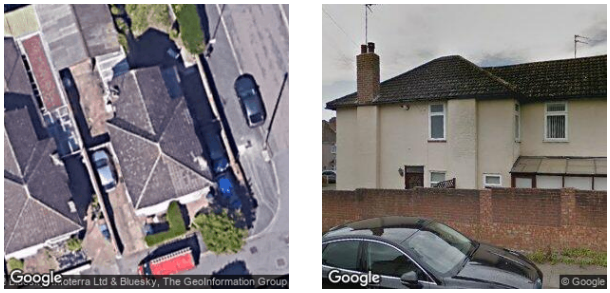
3. Dataset

The final dataset for our study consists of 39,605 buildings. Each building is represented by an aerial image, a street view image, satellite-borne heat loss measurements derived from land surface temperature (LST) data [11, 9], and OSM-derived footprint polygons [5]. Moreover, each building has an associated ground truth energy performance label which specifies a building's energy efficiency in terms of seven classes ranging from "A", the best, to "G", the worst. The ground truth energy efficiency labels have been obtained from [10].

The ensure that our dataset represents the real world as closely as possible, we have collected building-level observations from four cities and regions across the United Kingdom: 21,607 buildings from the city of Coventry, 6,834 buildings from Westminster in the city of London, 7,464 buildings from the city of Oxford, and 3,700 buildings from Peterborough. The cities for our case study have been selected based on the availability of the aforementioned data sources and their variety in urban landscapes. To measure how well our approach generalizes across cities in the same country, we split out dataset into a training set of 32,315 buildings which have been randomly sampled from Coventry, Westminster, and Oxford, and a validation set of 3,590 buildings consisting of the remaining buildings from the aforementioned regions. The test set consists of 3,700 buildings and is exclusively taken from the city of Peterborough.

It is important to note that the dataset exhibits a strong class imbalance. In the original dataset each building is assigned an energy performance label between "A" and "G", with 0.07% of the buildings belonging to class *A*, 2.68% to *B*, 16.83% to *C*, 46.36% to *D*, 26.29% to *E*, 5.85% to *F*, and 1.93% to the worst class *G*. A visual example for a building in Coventry is depicted in Figure 1. In this paper, we aim to model the energy efficiency of a given building in a binary fashion, i.e. differentiate between efficient and inefficient buildings. To do so, the buildings in our dataset are grouped according to their respective energy performance label. Buildings falling into the categories "A" to "D" are considered to be energy efficient (65.94% of the data), while buildings falling into the categories "E" to "G" are considered to be inefficient (34.06% of the data). The reason for this binary grouping is based on the idea that when applying the algorithm on a large-scale in the real world, we would like to reliably identify the buildings with the poorest energy efficiency, i.e. those which generally represent the highest retrofit potential. A binary grouping significantly simplifies the modeling process, adapts better to data-scarce scenarios, and reduces the class imbalance while providing meaningful insights into the existing building stock.

To obtain the final dataset, the addresses in [10] are geocoded and spatially joined with the OSM-based building footprints and the respective heat loss signal. In cases where a given building would consist of multiple flats and floors, only the label of the least efficient top floor apartment would be considered for the sake of our estimation. This is because the satellite-borne heat loss measurements derived from [9] mostly captures a building's heat loss through its roof.



(a) Aerial Image (b) Street View Image

Figure 1: Aerial and Street View Image for a Residential Building with EU Energy Label *E* in Coventry, UK.

3.1. Aerial and Street View Imagery

The aerial and street view images are obtained from [3]. For each building footprint, the coordinates of the respective centroid define the location for which we download the imagery. The street view images are downloaded with a field of view equal to 50 and the aerial images with a zoom level of 20.

3.2. Land Surface Temperature (LST) Data

The land surface temperature data which provides the heat loss information is obtained from [11] and builds upon [9] to obtain the Landsat-8-based long wave infrared product in an upsampled spatial resolution of 30x30m. The heat signal per building footprint is an average of all the building-specific LST observations for which the ground temperature at the time of data collection has been below a threshold of 5°C.

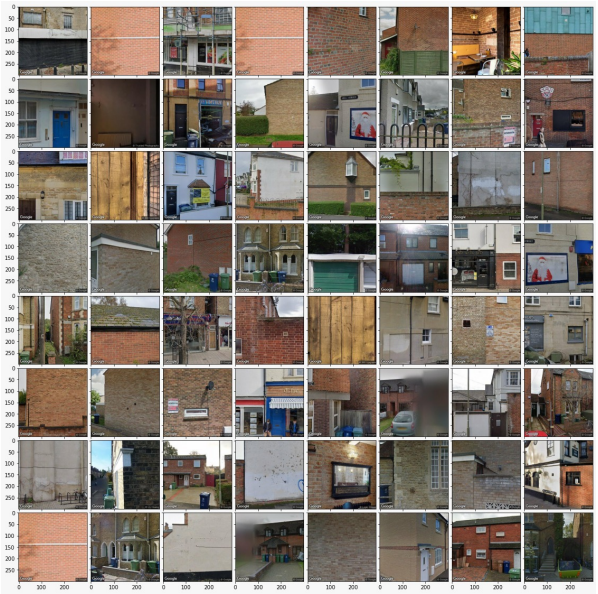
4. Methodology

While previous works have estimated building energy consumption from street view and aerial imagery [22], our work shifts the focus to building energy efficiency and also includes an entirely new data source in the from of building-level heat loss information derived from LST observations.

4.1. Data Cleaning with K-Means Clustering

Working with a variety of different data sources, maintaining a high dataset quality is crucial to build and train effective models. Unlike LST data obtained from [9] and

(a) Clean Street Views



(b) Noisy Street Views

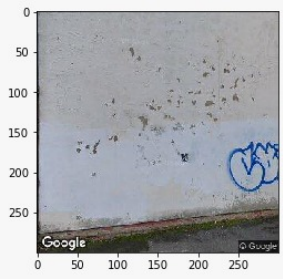
Figure 2: Clusters representing clean and noisy street views.

aerial images obtained from [3] which are pre-processed, street view images exhibit a high degree of variance and require significant data cleaning in order to filter out noisy examples.

Cleaning the street view images is a multi-step process. First, all street view images are encoded into an embedding space with an inception-v3-based feature encoder network [26]. Then, we conduct K-Means clustering on the embedded street view images in order to find images which do not depict meaningful features for our modeling task, i.e images taken indoors or which do not show a building facade. To increase the quality of the dataset, image clusters without meaningful information are removed from the final dataset. To verify the cluster-based decisions, we manually re-evaluate all removed street view images and add them back into the dataset if the cluster's assignment is erroneous.

Embedding and clustering the street view images with a pre-trained inception-v3-based network reveals interesting patterns in our dataset as depicted in Figure 2. In Figure 2, we juxtapose a cluster of street views depicting valid building images with a cluster of street views which mostly consists of brick walls and bears little to no signal with respect to building energy performance. The embedded and clustered street views also provide insights into different urban environments and architectural styles which in turn can be indicative of socio-economic status and construction dates, as depicted in Figure 4.

Based on the street view embeddings, we are also able to train a nearest neighbor model to perform a semantic search



(a) Reference Image

(b) Nearest Neighbors

Figure 3: A reference image and its nearest neighbors.

within the embedding space. This means that for a given image, we are able to automatically retrieve semantically similar images as illustrated in Figure 3.

4.2. Data Cleaning with Semantic Segmentation

Apart from cleaning the street view dataset with k-means clustering and semantic search, we also make use of a pre-trained semantic segmentation model from [18]. Being trained on the Cityscapes dataset [6] which has been developed for semantic urban scene understanding, particularly autonomous driving, the segmentation model can be used as an optional pre-processing step in order to remove the sky from street view images, as illustrated in Figure 5. The intuition behind removing the sky from street view images is based on the observation that the sky does not provide

any signal with respect to building energy performance but introduces a significant amount of noise through it's variability in terms of location, date, and time of observation. Based on [17], we decided to not mask out any other image features, such as cars, sidewalks, and vegetation, as these features can provide meaningful clues with respect to the spatial context and socio-economic status of a given property.

4.3. Baselines

For our baselines, we employ standard computer vision performance baselines that work directly on the raw data. The baseline models include k-nearest neighbor (k-NN) and support vector machine (SVM) models and are trained for binary classification. Both baseline models are compared to a third model which simply predicts the majority class for every data point (Majority Model). The rationale behind the choice of baselines is that both model types, SVMs and k-NNs, work well on high-dimensional data which is essential when working with image data input. To further reduce the dimensionality, memory requirements, and processing time, all our baseline models are only trained on street view imagery.

4.4. End-to-End Deep Learning Architecture

Apart from the baseline models described in the previous section, this study also proposes and evaluates an end-to-end deep learning architecture for predicting building energy efficiency from LST data, street view, and aerial imagery. To do so, street view images and aerial images are each encoded with an inception-v3-based feature encoder network [26]. These encoder networks have been pre-trained on ImageNet and their weights remain frozen. In contrast to the inception-v3-based classification network, the feature encoder network drops the final affine layer and maps each image into a 2048-dimensional embedding space. In the embedding space, building-specific heat loss information and the building's footprint area are added as additional feature dimensions and concatenated to the embedded street view and aerial image vectors. Lastly, a two-layer feed-forward neural network with a ReLU nonlinearity is trained on the joint embedding space in order to predict whether a given building feature vector is considered to be energy efficient (output is zero) or inefficient (output is one). By making use of a two-layer head, the model has the opportunity to non-linearly combine the street view, aerial, LST and footprint data as opposed to a one-layer head which could combine these features only in a linear fashion.

5. Results

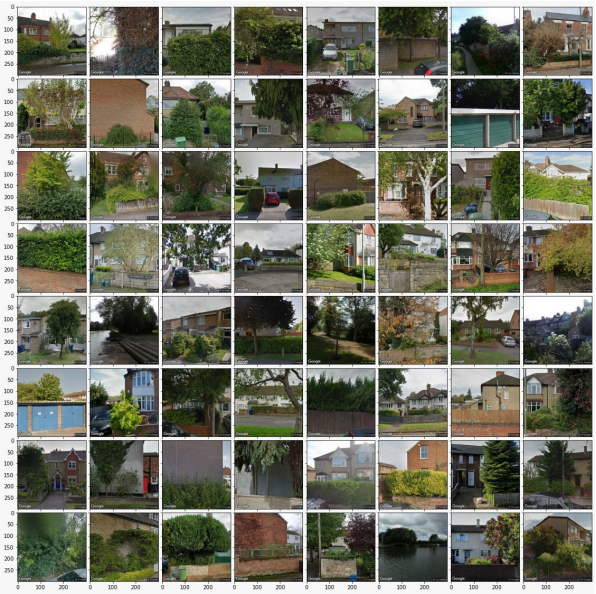
5.1. Evaluation Metrics

The evaluation metrics chosen for our binary classification task are the macro-averaged and weighted precision, recall, and F1-score. The rationale behind choosing macro-averaged scores is linked to the label imbalance in our dataset. To discourage the models from simply predicting the majority class (energy efficient), macro-averaged scores ensure that all classes are assigned equal weight during evaluation. This is especially important as the label distribution can vary significantly between different cities. This is because each city is characterized by its unique history and the buildings can significantly vary in terms of architectural style, age, and potentially materials, all of which influence the energy efficiency of the respective building stock. Thus, opting for macro-averaged evaluation metrics can improve the model's robustness against distribution shifts when deploying the model to new geographies.

5.1.1 Baselines

For our k-NN model, we re-scale all street view images to a size of 100 by 100 pixels and choose a number of neighbors $k = 3$ found via a hyperparameter search, tuned on the validation set. For the SVM baseline, the street view images are again re-scaled to 100 by 100 pixels. The SVM classifier uses a radial basis function kernel and an L_2-regularization term with an inverse regularization strength of $C = 1.0$, chosen via a hyperparameter search on a logistic scale ranging from $C = 1e^{-4}$ to $C = 1000$. For SVM models, the same kernel values are re-used over multiple training iterations and saved in a cache. If this case is not exhausted due to memory requirements, the training time scales on the order of $O(n_{\text{features}} \cdot n^2_{\text{samples}})$. However, once the cache is exhausted, the training time scales according to $O(n_{\text{features}} \cdot n^3_{\text{samples}})$ which is why we train on only the first 5 000 samples in the training set for a maximum of 10 000 iterations, given resource limitations.

As Table **??** shows, the SVM classifier outperforms both the k-NN model and Majority Model baselines across the macro-averaged precision, recall, and F1-scores. Notably, both the k-NN model and the SVM model perform significantly better than the Majority Model which speaks for the predictive potential of our data sources. However, even the SVM-based model cannot exceed a marco-average F1-score of 57%, leaving a lot of room for performance gains which can be achieved by more sophisticated modeling techniques. We hypothesize that the SVM model performs better in the binary classification setting due to its kernel-induced capability to effectively model data in high dimensions. The k-NN model, although being an extremely sim-



(a) Suburban Street Views

(b) Urban Street Views

Figure 4: Clusters representing suburban and urban street views.

(a) Original Street View

(b) Segmented Street View

Figure 5: Masking out the sky with semantic segmentation.

ple non-parametric algorithm, also performs significantly better than the Majority Model. This indicates that even pixel-by-pixel comparisons between street view images can be correlated with building energy efficiency. However, when validating the k-NN model and comparing its predictions on the Peterborough-based test set to the metrics obtained from the validation set, it becomes clear that the k-NN model is particularly susceptible to distribution shifts between cities, even within the United Kingdom.

5.1.2 End-to-End Deep Learning Model

The end-to-end deep learning model fuses the different data sources in a single model architecture and is thereby able to more flexibly combine the signals from the different inputs. As a result, the end-to-end deep learning model achieves a macro-averaged F1-score of 62.06% and outperforms the k-NN and SVM-based baseline models by 5.62 to 11.47 percentage points, respectively. The model hyperparameters are chosen via grid-search. The best-performing model is trained with an Adam optimizer, a batch size of 16, a learning rate equal to $7e^{-4}$, and class weights that are inversely proportional to the class frequency.

5.2. Qualitative Results

To increase the interpretability of the proposed end-to-end deep learning model, we build upon the integrated gradients attribution method described in [25]. To do so, we adopt the codebase in [16] to work with multiple data sources. This method is then used to quantify the joint and interdependent contributions of the street view and aerial images with respect to the final model's predictions. As a result, this method attempts to explain the model's logit score prediction of the energy efficient class relative to the logit scores of the energy inefficient class. Using Equation 1, we derive attribution maps for both street view images and aerial images to quantify the two image sources' impact on the difference in logit scores between the energy efficient and inefficient class.

$$\text{Attr}(x_i) = (x_i - x') \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \cdot (x_i - x'))}{\partial x_i} d\alpha, \quad (1)$$

Table 1: Quantitative Model Results on Peterborough Test Set (in %)

Model	Binary Classification (Macro)		
	Precision	Recall	F1
Majority Model	40.55	50.00	44.78
k-Nearest Neighbor (k-NN)	51.97	51.14	50.59
Support-Vector Machine (SVM)	56.59	56.31	56.44
End-to-End Deep Learning	**61.99**	**62.86**	**62.06**

x_i represents the image the attribution map is computed for, F is the difference in the model's prediction of the efficient class logit score and the inefficient class logit score, and x' is a baseline image with respect to which the attribution map is generated. In our case, the baseline is simply an image with random pixels. We take the integral in Equation 1 separately with respect to the street view image and the aerial image in order to generate an attribution map for each of the end-to-end model's image inputs. Since the definite integral in Equation 1 is difficult to compute, it is approximated by a number of discrete intervals as shown in Equation 2 where we choose $m = 50$.

$$\text{Attr}(x_i) = (x_i - x') \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \cdot (x_i - x'))}{\partial x_i} \quad (2)$$

Finally, the integrated-gradients attribution maps were smoothed using a Gaussian blur filter. Figure 6 depicts an example of a smoothed street view and aerial attribution map computed for a residential building located in Peterborough, UK, which is part of our test set. The street view attribution map reveals that the model puts significant weight on the building on the left-hand side of the image as well as on the tree in the garden, two of the windows and part of the roof and solar panels. On the aerial attribution map, the model appears to focus on various parts of the image, particularly focusing on the building's roof and driveway. Given the qualitative results obtained from our application of the integrated gradients method, we cannot draw definite conclusions about whether our end-to-end deep learning model effectively focuses on relevant parts of the street view and aerial images. Overall, the qualitative analysis of our results suggests that the model tends to place emphasis on the buildings itself and not surrounding irrelevant features such as the sky or vegetation. In addition, this observation appears to be consistent with joint attribution maps of other examples in the test dataset. However, the visualized results indicate that the end-to-end deep learning model does not yet manage to identify building energy efficiency related image features consistently.

5.3. Ablation Study

In order to obtain a better understanding of the predictive power of each data source, we conduct an ablation study in which models trained on a variety of different combinations of data sources are evaluated against each other. In other words, we evaluate models predicting building energy efficiency from only one data source, i.e. aerial imagery, street view imagery, land surface temperature data, or footprint area alone. Moreover, we also evaluate the performance of models which are trained on a combination of these features, i.e. LST and footprint area, for which results are shown in Table 2 .

The ablation study reveals that averaging the predicted probabilities of three separate models trained on street view data, aerial imagery, and LST plus building footprint data, respectively, achieves the best performance level. This indicates that the signals picked up in the different data sources are complementary and can be used in combination to better predict the energy efficiency of buildings. Notably, the averaging of the separate models' probabilities currently achieves higher scores across all macro-averaged evaluation metrics than the end-to-end deep learning model which learns how to combine the data sources itself. This suggests that the currently limited available amount of data in addition to model architecture choices for the end-to-end deep learning model are not sufficient to accurate learn how to combine the different data sources in an end-to-end fashion. Our results show the promise of combining different remotely-sensed data sources to predict building energy efficiency, with new means of combining relevant data sources in an end-to-end fashion being an important direction for future research.

6. Conclusion and Future Work

In summary, this work presents a novel way to estimate building energy efficiency from remotely sensed data sources only. As a result, our work makes three major contributions. To begin with, this is the first work which preprocesses, utilizes, and evaluates satellite-borne heat loss data to estimate building-level attributes such as building energy efficiency. Second, we create a novel building-level

(a) Original Street View (b) Street View Attribution Map (c) Original Aerial Image (d) Aerial Attribution Map

Figure 6: Joint-Effect Attribution Maps Using Integrated Gradients for a Residential Building in Peterborough, UK

Table 2: Ablation Study on Peterborough Test Set (in %)

Model	Binary Classification (Macro)		
	Precision	Recall	F1
End-to-End Deep Learning (2 Inception-v3 Models + 2 Linear Layers)	61.99	62.86	62.06
Inception-v3 (Street View) + Log. Regression	60.31	60.73	60.51
Inception-v3 (Aerial Imagery) + Log. Regression	59.07	58.99	59.03
Land-Surface Temperature + Building Footprint + Log. Regression	57.56	55.82	56.24
Averaging Model Predictions	**65.86**	**62.51**	**63.67**
Land-Surface Temperature + Log. Regression	57.93	55.71	56.15
Building Footprint + Log. Regression	57.71	56.01	56.45

dataset consisting of almost 40,000 buildings across four diverse regions in the United Kingdom. Third, we study and quantify the predictive power of each data source in an ablation study and compare a variety of baseline models to an end-to-end deep learning architecture.

Our results indicate that in the binary setting of predicting building energy efficiency the end-to-end deep learning model achieves a macro F1-Score of 62.06%. When comparing the performance of the end-to-end deep learning model with the baseline models and the models trained in the ablation study, we can conclude that the end-to-end deep learning model performs superior to all but one approach; the ablation study reveals that by averaging the model predictions, we are able to improve overall performance by more than 1.6 percentage points in terms of the macro F1-score. This indicates the usefulness of combining different remotely-sensed data sources to effectively predict the energy efficiency of buildings. However, in order to learn to combine these different data sources in an end-to-end fashion which outperforms the averaging of single-data-source models, we suggest to collect more data and conduct further research regarding the architecture choices of end-to-end deep learning models.

There are multiple ways beyond improving upon our end-to-end deep learning model in which future work could extend this study. Based on our results, further research could examine the potential of remotely sensed data to estimate household-level energy consumption. Moreover, other remotely sensed data modalities including Lidar-derived point clouds could be added and evaluated with respect to their predictive power in terms of building energy efficiency and consumption. Finally, future work could additionally explore how remotely-sensed data could be effectively complemented by other data sources to improve performance while ensuring scalability.

7. Contributions and Acknowledgements

Kevin Mayer: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Data curation, Writing.
Lukas Haas: Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Data curation, Writing.
Tianyuan Huang: Data Curation, Methodology.
Martin Fischer: Funding, Resources.

References

[1] J. Bin, B. Gardiner, E. Li, and Z. Liu. Multi-source urban data fusion for property value assessment: A case study in

philadelphia. *Neurocomputing*, 404:70–83, 9 2020. 2

[2] J. S. Chou and D. S. Tran. Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. *Energy*, 165:709–726, 12 2018. 2

[3] G. Cloud. Google cloud platform. 3, 4

[4] E. Commission. Energy efficiency in buildings. 2020. 1

[5] O. contributors. Openstreetmap, 2017. 3

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, D. A. Rd, and T. U. Darmstadt. The cityscapes dataset for semantic urban scene understanding. 4

[7] C. Deb, Z. Dai, and A. Schlueter. A machine learning-based framework for cost-optimal building retrofit. *Applied Energy*, 294, 7 2021. 2

[8] T. R. Dougherty, T. Huang, Y. Chen, R. K. Jain, and R. Rajagopal. Schmear: Scalable construction of holistic models for energy analysis from rooftops. pages 111–120. Association for Computing Machinery, Inc, 11 2021. 2

[9] S. L. Ermida, P. Soares, V. Mantas, F. M. Göttsche, and I. F. Trigo. Google earth engine open-source code for land surface temperature estimation from the landsat series. *Remote Sensing*, 12, 5 2020. 3

[10] H. . C. D. for Levelling Up. Energy performance of buildings data: England and wales, 2021. 3

[11] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 12 2017. 3

[12] E. J. Hoffmann, Y. Wang, M. Werner, J. Kang, and X. X. Zhu. Model fusion for building type classification from aerial and street view images. *Remote Sensing*, 11, 6 2019. 2

[13] C. E. Kontokosta. Predicting building energy efficiency using new york city benchmarking data. 2012. 2

[14] S. Krapf, N. Kemmerzell, S. K. H. Uddin, M. H. Vázquez, F. Netzler, and M. Lienkamp. Towards scalable economic photovoltaic potential analysis using aerial images and deep learning. *Energies*, 14, 7 2021. 2

[15] S. Lee, S. Iyengar, M. Feng, P. Shenoy, and S. Maji. Deeproof: A data-driven approach for solar potential estimation using rooop imagery. pages 2105–2113. Association for Computing Machinery, 7 2019. 2

[16] K. Leino, Rshih32, D. Gopinath, , Anupam, Shayaks, MacKlinkachorn, and Caleblutru. truera/trulens: Trulens, 2021. 6

[17] X. Li, C. Zhang, W. Li, Y. A. Kuzovkina, and D. Weiner. Who lives in greener neighborhoods? the distribution of street greenery and its association with residents' socioeconomic conditions in hartford, connecticut, usa. *Urban Forestry and Urban Greening*, 14:751–759, 2015. 5

[18] Mapillary. inplace-abn github repository, 2022. 4

[19] K. Mayer, B. Rausch, M. L. Arlt, G. Gust, Z. Wang, D. Neumann, and R. Rajagopal. 3d-pv-locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3d. *Applied Energy*, 310, 3 2022. 2

[20] A. D. Pham, N. T. Ngo, T. T. H. Truong, N. T. Huynh, and N. S. Truong. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, 260, 7 2020. 2

[21] B. Rausch, K. Mayer, M.-L. Arlt, G. Gust, P. Staudt, C. Weinhardt, D. Neumann, and R. Rajagopal. An enriched automated pv registry: Combining image recognition and 3d building data. 12 2020. 2

[22] M. Rosenfelder, M. Wussow, G. Gust, R. Cremades, and D. Neumann. Predicting residential electricity consumption using aerial and street view images. *Applied Energy*, 301, 11 2021. 2, 3

[23] A. Streltsov, J. M. Malof, B. Huang, and K. Bradbury. Estimating residential building energy consumption using overhead imagery. *Applied Energy*, 280, 12 2020. 2

[24] M. Sun, C. Han, Q. Nie, J. Xu, F. Zhang, and Q. Zhao. Understanding building energy efficiency with administrative and emerging urban big data by deep 1 learning in glasgow 2 3 understanding building energy efficiency with administrative and emerging urban big. 2022. 2

[25] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. 6

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens. Rethinking the inception architecture for computer vision. 4, 5

[27] A. Tsanas and A. Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 6 2012. 2