**Project Report**

**On**

**Solar PV in Aerial Imagery**

By: Team 5

# **Table of Contents**

1. Abstract
2. Introduction
3. Background
4. Data
5. Methods
6. Results
7. Conclusion
8. Roles
9. References
10. **Abstract**

This should be the one paragraph that captures the significance of what you did and why you did it.

Several models to classify whether images of houses contain a solar panel or do not contain a solar panel. Two models were determined as effectively classifying the images into the correct categories, as compared to a basic logistic regression model. The first was a Histogram of Oriented Gradients Support Vector Machine, with an accuracy on the test set of 83%. The second model was a Computational Neural Network, with an accuracy of correctly classifying the test images of 96%. Together, these models provide an effective tool to allow for the classification of residential homes as either having or not having solar panels. These tools can be used to further understand the distribution of solar panels in residential areas, allowing governments or other agencies to examine the effectiveness of their policy proposals as related to alternative energy and climate change.

1. **Introduction**

Provide a description of the problem and the value in finding a solution, motivate your reader as to why he/she should care about this question. The idea is to get your reader excited about the solution you are about to present.

Global climate change is a serious problem with far reaching, long term consequences to all nations and economies. Most energy generation in the world comes from fossil fuels, the byproducts of which are implicated in causing the current phase of global warming. Much research has gone into looking for alternative sources of energy, including wind and solar power. However, these technologies still continue to be very expensive compared to more traditional means of energy generation, such as coal plants. To try and counteract this, many governments, including the government of the United States, have funnelled money into both alternative energy research and alternative energy production. Despite numerous government funds, these technologies still only constitute a minority of energy generation in the US. The ultimate goal of this project will be to help the government inform best policies to encourage a broader use of renewable energy sources, particularly as applied to residential consumers. To that end, this project will focus on residential consumer’s use of solar panels on their houses. This is an attractive research group because it can measure both popular support for renewable energy sources (as opposed to industries complying with federal law) and the effect of government policies to encourage the use of renewable energy sources. There have been numerous programs in several US states to try and encourage the use of solar panels, including <https://www.energysage.com/solar/cost-benefit/solar-incentives-and-rebates/>

This project seeks to use satellite imagery to computationally determine if a house has a solar panel. This can help governments to determine if their policies of financially supporting solar panels are yielding more solar panels on private homes. This process can then be used in conjunction with other studies to determine if government policies concerning residential solar panels are effective at increasing the use of solar panels.

1. **Background**

This section should cite problems that have been previously addressed that relate to your work, and the key takeaways of the studies that explored that work. The idea here is to place the problem you’re working on in context and to let the reader know that you’re not working in a knowledge vacuum. For finding relevant literature, a good starting point is Google Scholar.

There are already some previous studies applied to machine learning methods to identify solar panels from satellite images. For example, “The DeepSolar Project” conducted by Stanford University, “constructed a comprehensive high-fidelity solar deployment database for the contiguous U.S.”[1] The purpose of the project is to create a database for people who interested in how the solar panels are installed in the different area in the U.S. to conduct further analysis or research. The team applied the classification method based on Google Inception V3 to identify whether there is any solar panel in the piece of the image.[1] If an image is classified as there exist any solar panel, segmentation would be conducted to estimate the size of the solar panels. Though the purpose of the project is slightly different from our project, the goal is the same- detect solar panels from satellite images.

To identify solar panels from satellite images, image processing and image classification are the important methods in this project. Image processing is a subfield of signal processing, which uses computers to process digital images. This has been studied for decades since the 1950s according to Azriel Rosenfeld.[2] Since the digital images are represented using matrices, we can do scaling, color conversion, image enhancement, etc. to the image by adjusting the value in the matrices. Moreover, image processing can be used to filter out the information from the high dimension features of the images. According to D. Lu, “implementing feature extraction, and selecting suitable variables for input into a classification procedure are all important” [3] since we can make full use of the features and also reduce the dimension of the data. Some image processing methods such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) [4] are famous as tools to extract informative features from the original image. With these filtered features, we can apply supervised machine learning methods to conduct image classification with high accuracy compared to use original images.

Previous work

What is Image classification

Analogues

Include Citations\*\*

1. **Data**

The data primarily comprised of satellite imagery of rooftops with or without a solar photovoltaic array(s). However, images without any rooftops/houses with solar array(s) installed were also part of the dataset.

Moreover, the solar array(s) had different type of structural configuration (or the arrangement) and orientation across the images in the data. The images in the dataset had a resolution of 101 x 101 pixels. And, the data was provided in two sets, training, and testing data.

* 1. **Data Description**

The *Training* data comprised of 1500 labeled satellite images comprising of 505 images *with* solar photovoltaic array(s) on the rooftop, and 995 images *without* any solar photovoltaic array(s). *Figure 1* shows the distribution of images for each class i.e. solar and non-solar

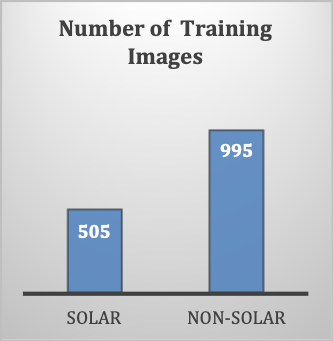


Figure 1: Number of Training Images for each label

The *Testing* data comprised of 558 unlabeled satellite image data *with* and *without* solar photovoltaic array(s).

*Figure 2* shows sample images *with* solar PV array(s) installed in different locations, configurations, and orientations having different intensity in color associated with the panels in the array. *Figure 3* shows sample images *without* solar PV array(s).



Figure : Examples of images with Solar PV array(s)



Figure : Examples of images without Solar PV array(s)

* 1. **Challenges associated with the data**

The primary challenge associated with image classification for this dataset was identifying certain edges and colors which were specific to solar array(s). Since the array(s) were mounted on the rooftop, which had similar edges to the array(s), distinguishing roof and arrays was the main challenge. Another interesting challenge with detecting the images was identifying array(s) which had different colour intensity and arrangement, distinguishing them from the images of similar structures such as top of car in similar colour, or roads and pathways. Another critical problem with image data is high dimensionality. Colored images with 101 x 101 pixels correspond to 30,603 features, which is considered as very high dimensional data, and will be computationally expensive when directly used in a machine learning algorithm. Dimensionality reduction is an important step when working with image data for faster computation.

One of the many traditional ways of solving the problem of feature extractions from these kinds of images is manipulation of the colour scheme to increase contrast between the object and noise. Contrast helps in distinguishing features (in other words the RGB or Grayscale value) of the object from the noise around it. Subsequently, dimensionality reduction methods such as Principal Component Analysis (PCA) could be used to reduce the dimensionality of the image features. Interestingly, resizing of images is a much faster way to reduce dimensionality, but it could lead to loss of important information from the image. Another method which is being widely used for dimensionality reduction and feature extraction with edge detection is Histogram of Oriented Gradients (HOG) which is discussed more in the coming sections.

Convolution is another widely use process for feature extraction and dimensionality reduction, primarily with conjunction with neural networks for image classification. The process of convolution is also discussed in the coming sections.

1. **Methods**

Present your machine learning solution including a description of your preprocessing, feature extraction, and classification techniques and why you made each of the choices you did. Discuss any methods that you didn’t create yourself and please cite relevant literature to support your claims. Also include a flow chart of your methodology to the reader can easily conceptualize your solution. Describe your approach to measuring generalization performance, what metric(s) you used and why. Write this section so that someone could recreate your results.

Three models were used which were eventually decided on as the final three models. The three final models were a logistic regression model (to set a baseline to test the other models against), a HOG SVM, and a CNN.

* 1. **Support Vector Classification using Histogram of Oriented Gradients (HOG) processed images**

Images were processed using Pillow library, followed by dimensionality reduction using Histogram of Oriented gradients method. Subsequently, the low dimensional data

* 1. **Convolutional Neural Network (CNN)**

Flow chart

Methods we did

Why we did

Motivation behind methods

Training and testing

Cross validation

ROC

1. **Results**

Include a complete performance assessment that includes your validation approach (cross validation, train/validate/test split, etc.) and the key metrics of performance for the problem (ROC curves, PR curves, confusion matrices if applicable, etc.). You should also compare your outcomes to at least one baseline model in addition to comparison against random chance guessing in the classification setting. At least one of your models must not be neural network based (although you are allowed to use neural networks for other models). Any comparisons of ROC plots or PR curves should be on the same plot so that they can be easily compared with one another. This section should be supported with multiple visualizations including examples where your method worked well, examples where it failed, and hypotheses supported by evidence as to why in each case.

1. **Conclusions**

It’s critical to have a strong ending and not just let the energy fizzle out of the report. Many readers, if pressed for time, will simply read your abstract and your conclusions. Very succinctly recap the problem you were studying and your approach to the solution. Focus on explaining the key takeaways from your work - these should not be merely a set of bullet points, but fleshed out conclusions. As you're writing your conclusions think about if the reader took nothing else away from reading your report, what would you want them to know most? Did you identify one particular approach that worked well? Was there a challenge that you faced that opens the door to working on solving a new problem? What avenues of research would you pursue next?

Out of the models created, the two best models, which achieved a significantly higher accuracy than a basic logistic regression, were the HOG SVM and the CNN models. The HOG SVM performed acceptably, achieving an accuracy of 83% on the test dataset. This model is much faster than the CNN, and could have applications where speed is necessary, but accuracy is not as much of a concern. In terms of strict accuracy, the CNN performed much better, with an accuracy of 96% at classifying images as correctly having or not having a solar panel on the house. The CNN clearly has an accuracy advantage over all other models at correctly classifying the images, however, it was much slower than any other model. This model would be preferred when accuracy is required, but speed is not a concern. As described in the introduction above, the solar panel image classification goal of this project could help inform policymakers or other stakeholders as they seek to experiment with and determine the best method to increase solar panel use. With an accuracy of 96%, the CNN model could be used to effectively determine the efficacy of policies targeted at increasing solar panel usage. This can be done by feeding satellite images into the model, allowing any level of government to analyze their policies effectiveness. Using this information, the government can then tailor the policies to further experiment with policies and legislation, with the ultimate goal of increasing solar panel usage, reducing reliance on fossil fuels, and combating climate change. Further research could include ways to boost the accuracy of this model even further, and to apply these image classification techniques to further questions of interest.

1. **Roles**

Since this is a team project, we want to know what your specific contribution was to this project. Provide detail on your individual role and how it contributed to the competition. Each team member should clearly articulate an individual role in multiple sentences making clear the team member's contribution.

Akshay used a Histogram of Oriented Gradients PCA, combined with a support vector machine model, to prepare the second model submitted for the final part of the competition. Akshay also wrote (parts of the report). Tzu Chun made several models, including (add more). Derek Wales made several models, including (add more). Melody Li created a Computational Neural Network model for the images, which was then submitted as part of the final part of the competition. Andrew Patterson created and tested several models while using the grayscale component reduction, including a test of a basic logistic regression and an experimental QDA test, both of which were submitted as part of the competition. Andrew also wrote the introduction, background, conclusion, references, and other parts of the rest of the report.

1. **References**

An alphabetical list of references cited in this work. A minimum of 10 are required. Consider using the Zotero citation manager for collecting and compiling your references. You should cite research papers, textbooks, etc. You do not need to cite common software packages like scikit-learn, etc., but if you do use a code package that is not common, include a link to the website you used.

[1] Jiafan Yu, Zhecheng Wang, Arun Majumdar, Ram Rajagopal. Stanford Magic Lab. Retrieved from: “<http://web.stanford.edu/group/deepsolar/home.html>”

[2] Azriel Rosenfeld 1969. Picture Processing by Computer, New York: Academic Press. Retrieved from: “<https://dl.acm.org/doi/abs/10.1145/356551.356554>”

[3] D. Lu 2005. A survey of image classification methods and techniques for improving classification performance Retrieved from: “<https://www.tandfonline.com/doi/full/10.1080/01431160600746456>”

[4] David G. Lowe 2004. Distinctive Image Features

from Scale-Invariant Keypoints. Retrieved from: “<https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf>”