RESEARCH ARTICLE

A Machine Learning Approach for Identifying Favorable Sites for Renewable Energy Installations

Phoenix Sheppard (student)[1 *]

1. High School for Mathematics, Science and Engineering, 240 Convent Ave, New York, NY, 10031, USA

* Corresponding author email: phoenixsheppard28@gmail.com

Abstract

This paper demonstrates the application of machine learning in determining suitability for utility scale sites for renewable energy production. Supervised algorithms such as Random Forest Classifier are employed in a Semi-Supervised learning process that allows underlying trends present in suitable and non suitable sites to be extrapolated to a mostly unlabeled dataset. The model iteratively trains from the pseudo labels it creates throughout this process, until all data points in the data set are labeled. This allows the small percentage of hand labeled data to be leveraged for use in the larger dataset. This open source tool can be used by anyone for the quick and precise determination of suitable locations for utility scale and personal installations based on the available renewable resources. It can also be used to influence future policy decisions around renewable energy.

Keywords

Robotics and Intelligent Machines; Machine Learning; Semi-supervised Learning; Renewable Energy; Solar; Wind.

Introduction

As the march into the 21st century continues inexorably, technological progress is expanding at an unprecedented rate. These new advancements often give us the feeling that we are living in a state-of-the-art world, but this could not be further from the truth. In fact, the majority of systems that power the world use technology that dates from decades or even centuries ago.[1-2] This fossil fuel technology causes many complications with the environment at various stages in the process. If an oil pipeline were to burst, for example, it could end up polluting the surrounding environment by leaking oil into the earth and water sources. These events destroy ecosystems and damage the planet's biodiversity, but they also are dangerous to human health. In December of 2022, this occurred with the Keystone oil pipeline in Kansas.[3] 14,000 barrels of oil were spilled into local water bodies after a stress fracture burst on the pipeline. This is an example of how fossil fuel energy can be harmful when done wrong, yet when it is done right the end result is always the same: pollution of the atmosphere and the rising of global temperature. Fossil fuels account for 25% of the greenhouse gas emissions in the United States, second only to the transportation sector at 28%.[4]

The solution on the horizon to this insurmountable problem is renewables. Renewables eliminate both issues with fossil fuels specified before, as they cannot pollute the environment by failing in transport, and the generation of renewable energy (RE) does not contribute to greenhouse gas emissions. In addition, with the advent of electric vehicles as of late, RE is poised to eliminate the two biggest causes of greenhouse gas emissions.

The most important factor to consider when expanding RE installations across the country is location. Location determines the amount of renewable resources– such as solar irradiance or wind speed– that can be harnessed by a RE installation. This directly correlates with the amount of power that can be generated by a RE installation. The current process of determining a RE installation site involves many different steps and variables to consider.[5] Note that this process involves much more than just the renewable resources at a location. One must also consider economic, population, and infrastructure

factors, among others. For example, the technical feasibility of a site is how well suited the site's physical and electrical infrastructure is. Another factor to consider is population density. A site cannot be too close to population centers, as they take up valuable land space in the case of photovoltaic installations, or make too much noise in the case of wind turbines. In addition policy considerations are one of the most important in determining a site's suitability. What unites all of these factors is that they are without a way to easily quantify them for use by a machine learning algorithm, because of the nature of the factors or the nuances within them, like population density.

This paper will be solely focused on determining suitability for land installations based on renewable resources, which is the most preliminary and easily quantifiable step in the entire process. The method of determining suitability used in this paper can also be applied for smaller scale RE installations like for personal or small town use, rather than exclusively large energy corporations. This is possible because of the limitation of only considering renewable resources in determining suitability, meaning a site determined as suitable or not suitable by the model would apply to any size installation at that location.

Note, this would only be possible if the tool was scoped to determine land based installation suitability, hence why this paper will not be covering the offshore renewable sector. By leaving the other steps of determining the suitability of a site that succeed this preliminary step of renewable resources- like economic, population, infrastructure, and policy concerns- to the users of this tool, I introduce many degrees of freedom to the applications of this tool.

This paper demonstrates a way of quickly and accurately filtering possible sites by their available renewable resources using a machine learning model. This can drastically reduce the total time in determining a site's suitability, and can allow for a large number of sites to be examined at once, without the need to scrutinize them one by one with a team of RE experts. This is achieved by teaching the model to see sites already determined as worthy for a utility scale (at least 10 MW capacity) RE installation, based on renewable resources, as suitable. In a way, the model can pick up on the patterns previously identified by experts and use them for new predictions. The model's predictions may also reveal new correlations previously unknown to experts.

To current knowledge, there is no study that has applied semi-supervised learning techniques to create a tool with the ability to identify suitable RE sites based on renewable resources and other weather variables. This tool is also unique in that it does not require expert interpretation of its output, however, it can be used at an advanced level to identify patterns between suitable and unsuitable sites. This is key if we are to make the switch to renewables.

Methods

This tool will be scoped for predicting suitability for solar and wind power installations because they are the two renewables that are the least dependent on terrain factors. Hydropower installations, for example, require detailed analysis about the water cycle in that area, and are hard to quantify.[6] Wind and solar power generation, oppositely, is directly affected by simply quantifiable weather variables that are less dependent on the immediate terrain, making them the perfect candidates for using machine learning analysis. In addition, two separate machine learning models are created - one for wind, and another for solar- because this reduces the complexity with labeling. The criteria for a suitable wind farm location differs from that of a solar farm location. Also, the hand labeling process detailed in the data preparation section only works when this condition is met. Therefore, It makes sense to compartmentalize these two separate predictive functions within two separate models.

This paper employs classification algorithms as the method to determine suitability of a location. The labels predicted are 0 for not suitable and 1 for suitable. Drawing from the popular machine learning and data science platform Kaggle, I see that Decision Trees and Random Forest Classifiers came in second place for most popular by use. Gradient Boosting techniques- developed by Jerome Friedman- were also ranked fourth most popular in the list.[7-8] What unites these different algorithms is that they all fall under the umbrella of ensemble methods, including parallel and sequential algorithms. Various reputable algorithms based on the survey were evaluated to find the best for this use case, which will be discussed more in the experiment section. Random Forest Classifier was determined to be the best suited algorithm for both the solar and wind models. This paper employs a semi-supervised training method illustrated in Figure 1.

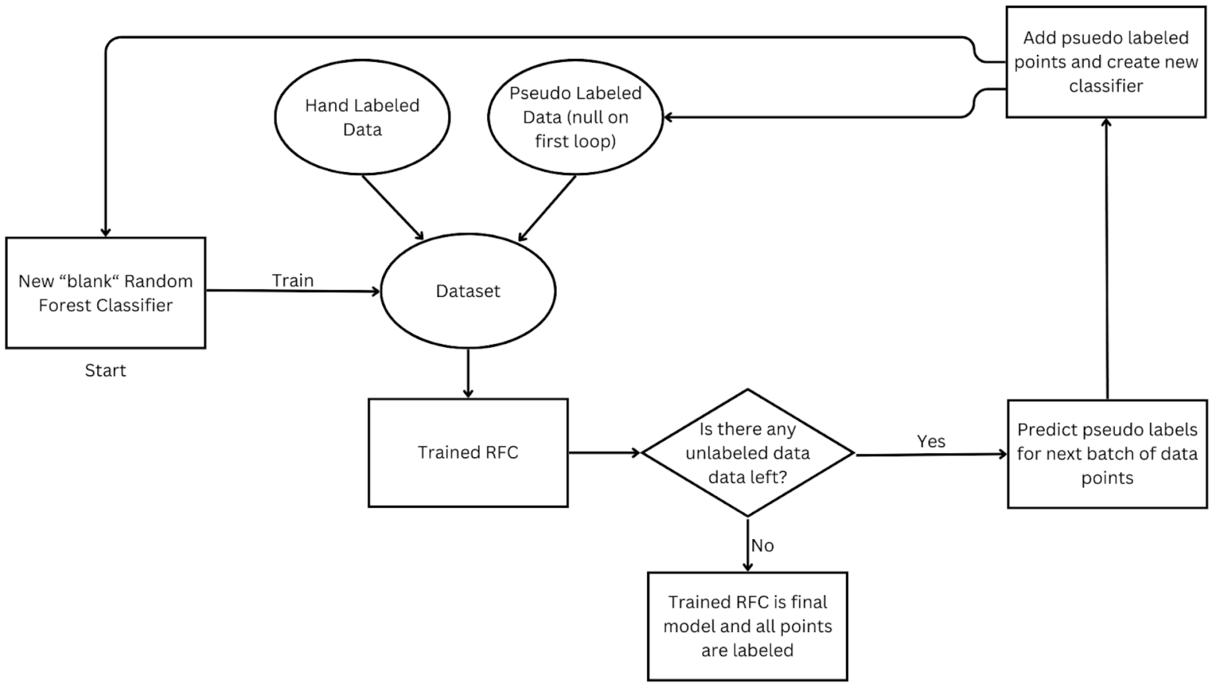


Figure 1. Flow chart for semi-supervised training process.

I chose to instantiate a new model at the beginning of each iteration of the process to combat a common problem in semi supervised learning methods: self reinforcement. If I had kept the same model and weightings throughout the process of adding new points to the dataset and training, it would have resulted in the model learning from its own pseudo labels primarily, regardless of whether they contained a propagated relationship from the initial high quality data. By keeping the size of each iteration limited to 100 new data points, I limited the scope of where this issue could have arised. At the end of all the iterations, a fully trained model is available along with predicted labels for the initially unlabeled part of the data set. In future applications of this tool, this same process can be used to build upon the trained model that I have created, or-if the user so desires- to start from scratch with a different set of initial seeded data and unlabeled data.

By applying the high-confidence technique used by Sohn et al, [9] I can ensure that only pseudo labels with high confidence level predictions are trained upon and used as ground truth. This is a highly effective form of consistency regulation, and has been used and studied many times in literature.[10-11-12] I will note that the literature referenced uses this technique within the context of computer vision, however, the reasoning behind using this technique still applies to classification models as I have used them.

Since we must use the pseudo labels for training, this method ensures that the error rate of wrong pseudo labels in the dataset is minimized. This is one of the most crucial parts of achieving a semi-supervised model that gives accurate predictions that users can then act on with confidence. After much experimenting, I determined that a threshold of 60% confidence in the prediction was the right balance between confidence and practicality for the pseudo labels. I confirmed that this confidence threshold was ideal by examining its final predictions for various points within the entire dataset. I found that only a small percentage of points were left with no label due to low certainty. I could confirm the predictions of many of the points that were labeled through logical reasoning and government bureau opinion such as the NOAA.

Data Preparation

This section will detail the data collection process and the logic behind the data collection. One of the requirements of the dataset was that it had to have a multitude of weather variables, at minimum including wind speed and solar irradiance, as they are the variables that have direct impact on the energy generation by wind turbines and photovoltaic installations respectively. It also had to include locations across the United States, in order to be able to accurately generalize based on what it had seen. I chose to use the National Renewable Energy Laboratory's National Solar Radiation Database. From that database, I chose to use the Physical Solar Model Version 3 (PSM) API, a synthetic weather model that itself derives various weather variables from its own calculations and other weather models.[13] With this dataset, I collected a number of weather variables in the 20 largest (by population) cities in each state excluding Alaska because of the range limitation of the model, totaling 980 cities. As part of the data cleaning process, unnecessary variables were removed if they conveyed the same information as other, more primary variables and measurements. This brought the total variable count from 14 down to 7 key variables that would then be put through further statistical feature selection.

I decided to use a modified Typical Meteorological Year (TMY) to represent the data of one location. The TMY was developed by the Sandia National Laboratories in New Mexico for the specific purpose of solar energy system studies, but it also includes metrics such as wind velocity.[14] It has since been built upon in many ways with TMY2 TMY3 and TMYx.[15] The TMY is widely used in climate and RE research and insight.[16-17-18] Generally, the TMY has been used for the comparison of different locations based on typical weather conditions over a decadal time span.[19] However, as a result of the recently accelerated effects of climate change, which can be seen in shifts in global and US temperatures, the weather patterns recorded across decades before hold less significance in the present more than ever.[20] As a result, I elected to use the year 2020 as the TMY for my data. As stated before, this allows me to easily compare locations based on recent weather data. It also has the effect of heavily optimizing the calculation time of the model.[21] The use of the TMY year by the NOAA[18] demonstrates averaging weather variables to gain deeper insight on the normal conditions of a location, and to identify larger trends across seasons and months. Similarly, I applied this approach to using the TMY, but also included irradiance and Wind Speed variables. This specific fusion of renewable resources and the TMY is not common, however, the underlying basis for averaging weather variables still applies to these renewable resources.

The next step of the data preparation process was to determine the high quality hand labeled data. As stated previously, there would be two models for solar and wind. Therefore, the 980 unlabeled data points from cities in the US would be copied and made into two different datasets to be built upon separately. To find sites that were extremely suitable for wind and solar installations, I decided to take the 120 largest (by MW) utility scale solar and wind installations in the US and label them 1 for suitable. Because utility scale RE installations are extremely costly to construct and maintain, the team determining the location for it needs to make sure that it has the best amount of renewable resources to make the best return on investment possible. This means that the locations where utility scale installations are placed are extremely suitable. Conversely, I also determined 120 unsuitable locations by taking 6 states from the original 980 data points and labeling the 20 cities within that state as 0 for unsuitable. I determined the 6 states for each dataset by looking at various factors, including at the wind speeds and irradiance in that state in comparison to others. This was the most important factor. I also took into account the amount of MW generated from solar and wind energy in that state, while also accounting for economic factors.[22-23]

By labeling the extremes of suitability by hand, the models would be able to clearly determine suitability by having stark contrasts appear in the original training set. The intention of this was to create a clear pattern of variable values that correlated to suitability and unsuitability within the dataset that the model could then build off upon as it trained and turned itself to make harder and harder predictions. The end result was 240 labeled data points and 860 unlabeled data points for the both datasets, leaving around 21% of the data labeled and the rest unlabeled. This sets the stage for semi-supervised learning to be used which is detailed in the method.

Feature Selection

To determine which of the 7 variables would be best used for prediction of suitability, I used the chi-squared score to rank the variables by most impact on classification label on the 240 seeded data points for both data sets (higher score is better). I can use the chi-squared score in this ranking fashion because each of the variables have the same degree of freedom and are in the same system/dataset. This eliminates the need for using the p value and null hypothesis. Table 1 displays the scores for the wind and solar dataset. They are both listed in descending order of score. It may be tempting to think that wind speed and GHI are the only variables that matter, because they directly affect the renewable power generation. However, using other seemingly unimportant variables can clue the model into what types of locations are suited for solar and wind power. This increases the model's predicting accuracy.

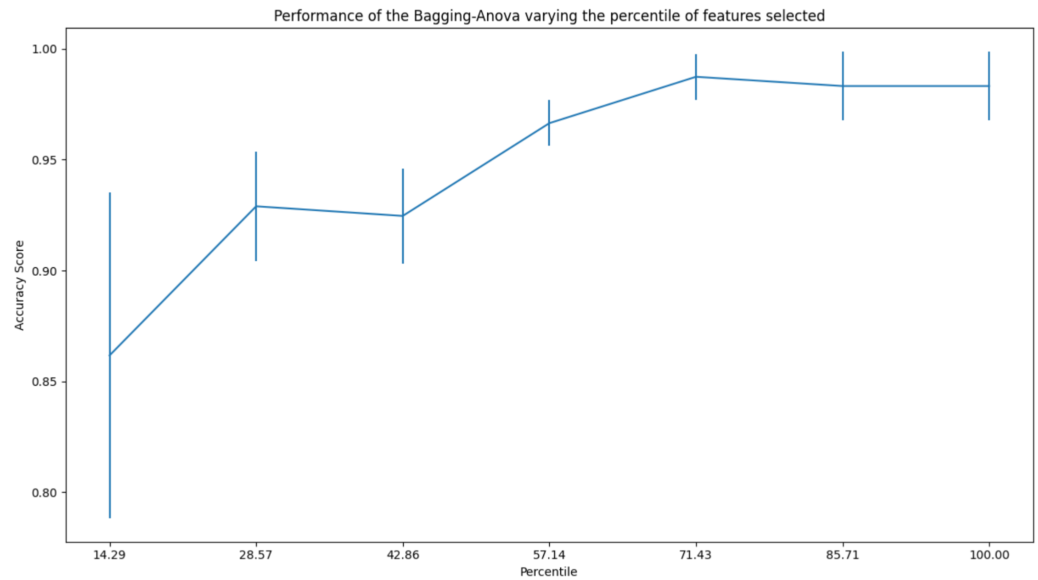
Wind Features Ranking		Solar Features Ranking	
Input Features	Chi2 Score	Input Features	Chi2 Score
Relative Humidity	308.202550	GHI	496.907810
Pressure	288.088902	Relative Humidity	162.636706
GHI	107.878059	Pressure	151.303983
Wind Speed	63.712324	Temperature	148.85482
Temperature	48.878754	Wind Speed	21.977538
Precipitable Water	29.019984	Surface Albedo	0.812015
Surface Albedo	2.797546	Precipitable Water	0.178297

Table 1. Ranking of wind and solar features by chi-squared scores.

The solar scores are to be expected, with GHI being by far the most important variable when predicting suitability. The relative humidity, pressure and temperature are at a much lower score than GHI, but are all generally clustered together, placing similar weights on their importance. After these variables, a significant drop off is observed at wind speed and especially surface albedo and precipitable water.

The wind scores, however, are less easily interpreted. Within the top 4 scores, not only is wind speed the least important variable, but is topped by the GHI. For the wind dataset, it would be expected that wind speed would easily top the list, but this discrepancy can be explained. Many of the suitable wind farms that were used in the dataset are in locations where the solar irradiance is very high, such as California, Texas, or New Mexico. Therefore, the importance of GHI here would be inflated and not representative of what it actually is. This explains why the GHI is so high on the list.

To verify the merits of the high scoring variables in the solar list and identify the truly important wind list variables we will use another method. By giving a machine learning model access to the seeded data points while varying the amount of features it can use to make predictions, we can find the ideal set of features to use. In other words, by comparing accuracy metrics from each distinct feature set, we can quantitatively determine the ideal set of variables to use for the solar and wind models. I chose to use the bagging classifier because of the random subsets of data it uses to train each tree in the ensemble.[24] This, combined with k cross-fold validation would ensure that the accuracy metrics recorded would not be affected as heavily by overfitting. I also chose to add another feature to the set to be considered each iteration based on their chi-squared score. For example, GHI scored first in the solar list, so it would be the first feature to be considered and would be by itself. The next iteration would add relative humidity (because it had the second highest score) to the features to consider, meaning GHI and relative humidity, and then so on until all the features were used. Figure 2 shows the accuracy as this process takes place for the wind and solar sets. Note that each percentile tick on the X axis represents a new feature from the list added to the group of features considered at that specific tick. The vertical bar represents the standard deviation of the cross fold validations while the middle of the bar represents the mean.



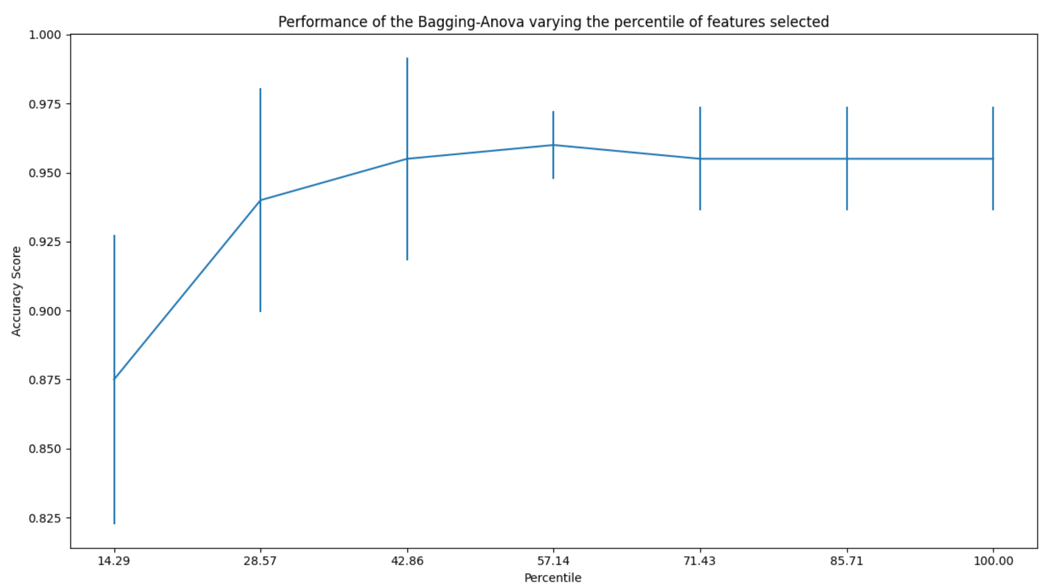


Figure 2. Classifier performance varying wind (top) and solar (bottom) features.

The wind graph shows an interesting development. As the first two features (relative humidity and pressure) are added, the accuracy increases. However, when the GHI is added at the third tick, the accuracy actually decreases. This signifies that GHI is detrimental to the classification accuracy and thus should not be considered in the variables to use for the wind model. The next two ticks after GHI is added are wind speed and temperature which also steadily increase the accuracy and shrink the standard deviation. Again, precipitable water and surface albedo do not make much of a difference on the classification strength. Therefore the final features to consider for the wind model are relative humidity, pressure, wind speed, and temperature.

As is shown in the solar graph, the mean accuracy peaks and the standard deviation reaches its lowest at the point when 4 features are included, meaning GHI, relative humidity, pressure, and temperature. This graph makes it fairly clear that these 4 features should be the final set to use when training the solar model. Wind speed, precipitable water, and surface albedo do not have much of an effect on the classification accuracy and thus can be dropped.

Model Selection

As stated before, the Kaggle survey top rankings for classification algorithms included all ensemble algorithms and decision trees. As such, I decided to test Random Forest Classifier (RFC), Bagging, Gradient boosting, and a Decision Tree to see which model would be the best suited for use on the wind and solar datasets. When scoring each model's performance I decided to use the F1-Score as a more complete measure of precision and recall beyond simple accuracy. I again used 10 cross folds to make sure that the high score of a model would not be due to overfitting. Figure 3 shows the bar graphs for the model selection process. The height of the bar graph represents the mean f1 score across the cross folds, and the number at the top is the standard deviation. Note that these models were evaluated with their default parameters outlined in the scikit-learn documentation.[25-26]

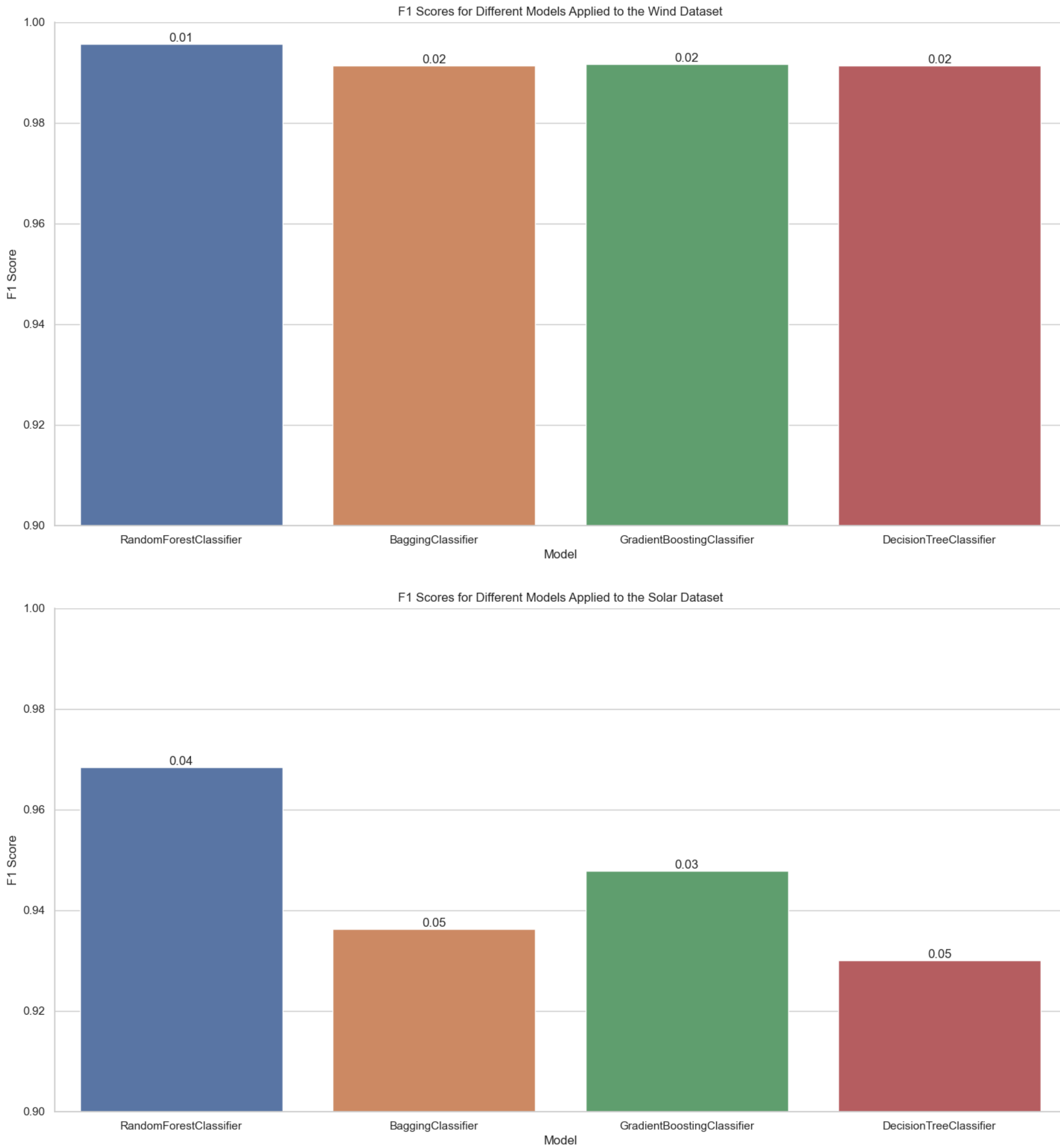


Figure 3. Wind (top) and solar (bottom) model selection.

For the solar graph, it is clear that the RFC performed the best out of all of the models tested, and with only a moderate standard deviation comparatively. For the wind graph, the bars are very close, but I decided to go with the RFC for its slightly higher accuracy and extremely low standard deviation. As such, I applied the RFC for both datasets. This is very helpful because tree based models do not require normalization of the data, and this is because they use the gini and entropy to split the internal nodes, not the calculated distance between features.[27] For both datasets there are variables with vastly different units and quantities, like pressure in hPa (hectopascals) and wind speed in meters/second, but by leaving these variables un-normalized, we can easily visualize the trees and understand the criteria the RFC uses for determining suitability. This will be detailed further in the results section

Hyperparameter Tuning

Now that I have determined RFC to be the ideal model, hyperparameter tuning will be done to determine the best settings of the model that would increase the accuracy while reducing overfitting. Note that we are using sci kit-learn's implementation of the RFC.[28] There are many different parameters that can be modified, but the ones I focused on were the ones that modified the structure of the decision trees within the ensemble, or how many trees would be generated. I kept the criterion as gini because it is been shown to have competitive accuracies with entropy (so long as the number of classes are low) while also considerably speeding up calculation times.[29-30] This is an important aspect to consider because people may be using this tool on many points to find the ideal location for a utility scale installation. Therefore, I will be hyperparameter tuning num estimators, min samples split, min samples leaf, max depth, and max leaf nodes.

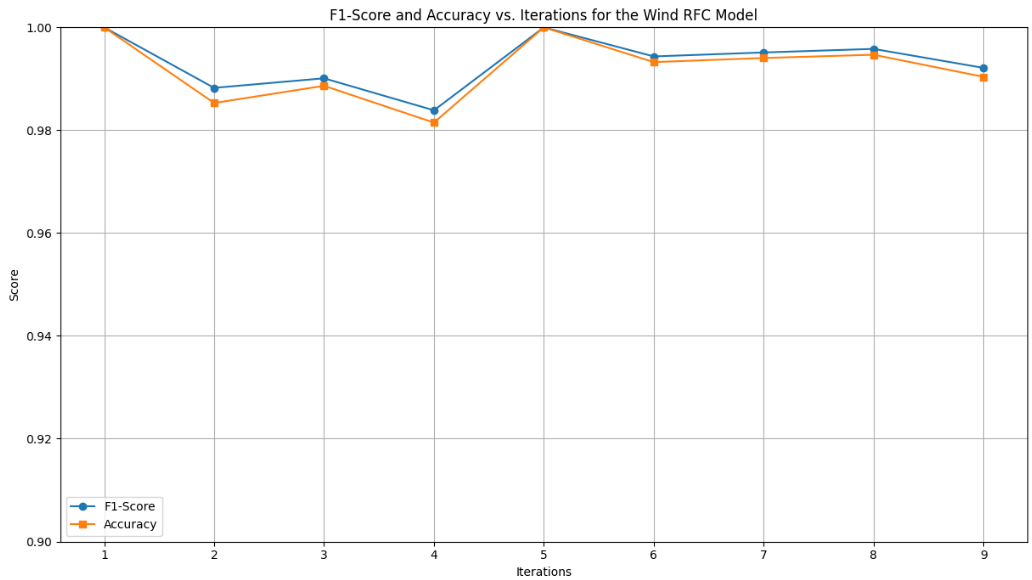
I utilized random search to randomly choose a value from a numeric range for each of the 5 parameters. After choosing the values, a model would be trained on the portion of the seeded data and evaluated on the rest using the f1-score. Again, 10 cross folds in the data were used here to make sure the parameters were not overfitting. Table 2 shows the ideal parameters for the wind and solar RFCs. The F1-score obtained with the following wind parameters was around 0.9917. For the following solar parameters the score was around 0.9648 . Both of these scores were rounded to the nearest ten thousandths. Considering the highest possible F1-score is 1, both these scores are phenomenal.

Wind parameters		Solar Parameters	
Parameter	Optimal Value	Parameter	Optimal Value
Num estimators	300	Num estimators	900
Min samples split	4	Min samples split	3
Min samples leaf	3	Min samples leaf	1
Max leaf nodes	27	Max leaf nodes	30
Max depth	5	Max depth	5

Table 2. Optimal wind and solar parameters.

Results and Discussion

This section will detail the findings of the experiment, including interesting labeling patterns, and metrics. Figure 4 displays the F1-score and accuracy at each iteration of the training process of the models. Table 3 displays the standard deviation of f1 score and accuracy across the iterations for the models. Note that every iteration beyond the first had pseudo labeled data within the pool to train off of. Also, it is important to remember that each time the fitting and predicting of the model was run, the results differed slightly each time, due to the inherent randomness of the algorithms. I selected graphs and subsequent metrics that I believe are most representative of the general trend for each model.



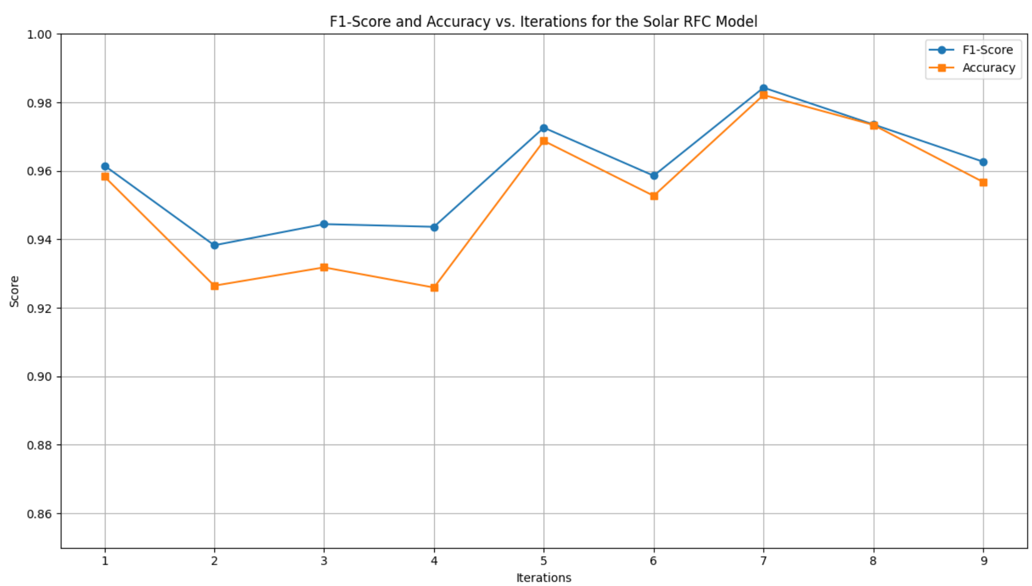


Figure 4. Wind (top) and solar (bottom) model metrics over iterations.

	Wind model		Solar model metrics	
Metric	F1-Score	Accuracy	F1-Score	Accuracy
Std	0.0050190	0.0058732	0.014265	0.019555

Mean	0.99329	0.99197	0.95995	0.95292

Table 3. Standard deviation and mean of F1-score and accuracy over iterations (rounded to 5 significant figures)

Clearly the wind model's predictive power is accurate. It achieved an exceptionally high mean score in both metrics and with a low standard deviation: less than 1 percent. This goes to show that the features determined in the feature selection process were valid and did effectively contribute to the model's predictive power.

While the solar model's predictive accuracy and F1-score is not as high as the wind model's, both metrics still manage to have a high score of around 95%. This could be for a number of reasons, but I expect that it is due to a fair number of the seeded suitable solar locations being located in California, which may have biased the model to predict suitability based on weather factors similar to California's. This will be examined in the discussion section further. Overall, both models managed to achieve a satisfactory level of robustness and accuracy.

As specified in the method section, precautions were taken so as to not have inaccurate pseudo labels infiltrate the dataset as ground truth labels. These measures were effective. For example, data points that would often flip labels on different runs of the model were able to be removed with the prediction confidence check. This is because an inconsistent label indicates an unconfident prediction that is solely dependent on the inherent randomness within each run of the model, rather than an identified pattern within the features of the variables of the model. Overall, this method has been able to keep the predictive accuracy of the model high, therefore producing labels that are truly representative of a site's suitability. For the data points with low confidence predictions, a team of experts could save them for human analysis like is done currently with many sites. This ensures that no possibly suitable site will go unnoticed by the users of this tool.

Now this section will cover interesting labeling patterns. Looking at the predictions for the Texas cities, for example, I can see that the northern cities like Fort Worth are predicted as suitable, while the southern ones like Houston are predicted as not suitable. This coincides with the wind speed map as shown in Figure 5.[31] This example also is a testament to the sensitivity of the model, as it is able to make differing predictions based on different weather variable values for each location, rather than generalizing about the suitability of a state as a whole.

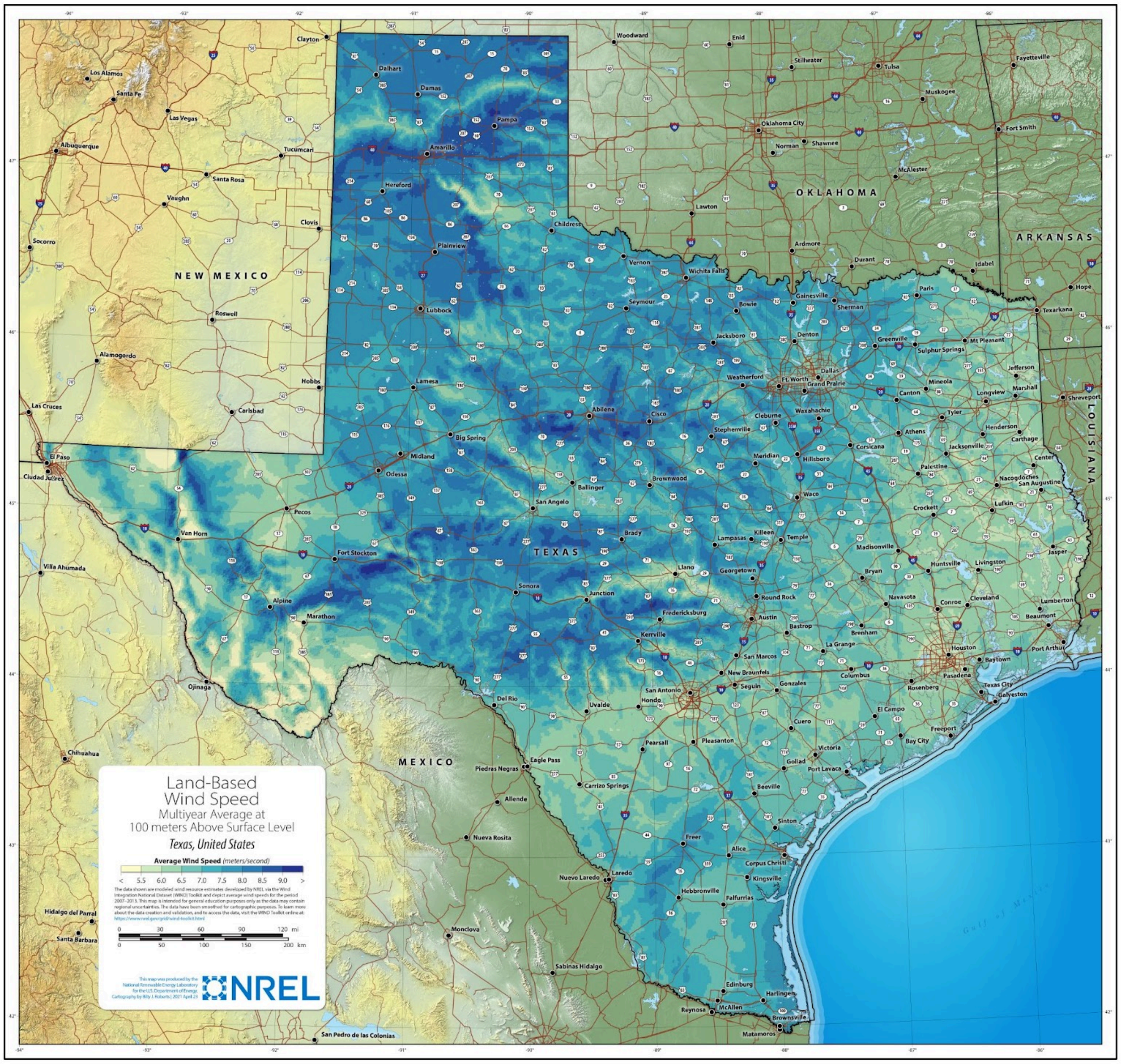


Figure 5. Average wind speed map of Texas, showcasing vast differences across the singular state.

For the solar model I performed a similar examination and found that it was also able to make different predictions within state borders. For example, Boise, Idaho vs Moscow, Idaho. Figure 6 shows a map of the Daily average GHI across the USA, and Figure 7 is that same map zoomed in on Idaho.[32] Boise is located in the 4.75-5 KWH/M^2/Day bracket (dark orange) while Moscow is located in the < 4 KWH/M^2/Day bracket (cream colored). The model labeled Boise as suitable and Moscow as not suitable, showing its sensitivity.

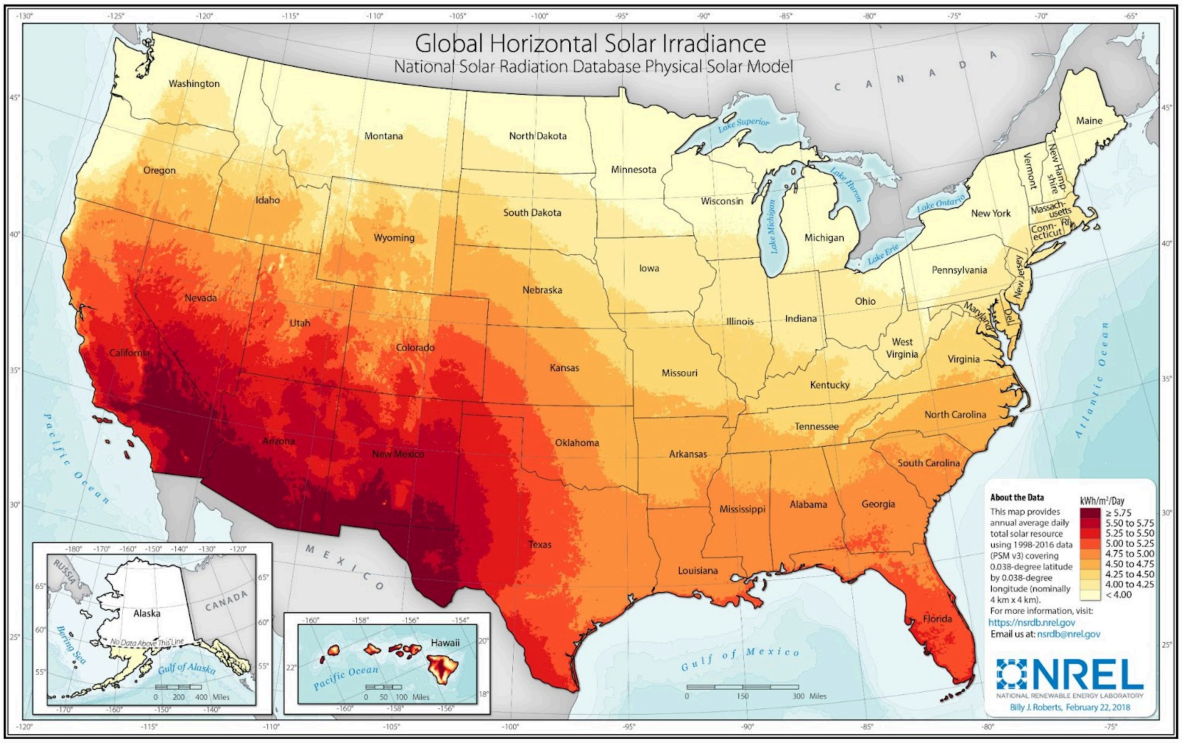


Figure 6. Daily GHI map of the USA.

Also, there is anecdotal evidence to support the sensitivity of the solar model claim. When I initially showed my mentor the predictions of the solar model, he was curious and wanted to know what the model predicted for the town he lived in. After inputting the weather data into the model for his hometown, it returned 0 for not suitable. This is interesting because according to him when people from the solar industry came to evaluate the location for possible home solar installations they came to the same conclusion: not suitable. Granted, this is an anecdotal example but it shows that the model can be used to circumvent the process of solar industry professionals coming to take measurements and physically evaluate a location. When considering this, it would not be a stretch to say that the wind model could be used in similar applications.

Conclusion

This paper has showcased a method of applying semi-supervised machine learning methodologies to a high quality labeled dataset in order to obtain highly accurate suitability labels for RE installations. In order to make this tool accessible to all people looking to apply this technology- homeowners, policy makers, corporate workers- I will make public the code and dataset that one would need to use this tool on GitHub, along with the instructions on how to apply this process to any new locations the user may desire. It will be located at https://github.com/phoenixsheppard28.

In order to resolve the climate crisis we as a human race have brought upon ourselves, we all must make conscious changes in our societies. No more can we rely on the deadly fossil fuels that further corrupt our world every day. Instead, with this technology, we must make the daunting but hopeful jump to renewables. Governments, corporations, and individuals can all play their part in some way or another. Together, we can all make the shift towards reclaiming our earth.

Acknowledgements

A very big thank you to Dr. Xiao Dong. He assisted me so much in this process of researching and writing and always encouraged me to be rigorous and thorough in my work. Thank you, Xiao.

References

1. Petros'yants, A. M. A pioneer of nuclear power, 1989.
2. Nuvolari, A.; Verspagen, B.; von Tunzelmann, N. The Early Diffusion of the Steam Engine in Britain, 1700–1800: A Reappraisal. *Cliometrica* **2011**, *5* (3), 291–321. DOI:10.1007/s11698-011-0063-6.
3. Kumar, A.; Nickel, R. *Reuters*. February 2023.
4. Sources of Greenhouse Gas Emissions | US EPA. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions (accessed 2023-10-24).
5. Site Considerations | US EPA. https://www.epa.gov/green-power-markets/site-considerations (accessed 2023-10-24).
6. How Hydropower Works. https://www.energy.gov/eere/water/how-hydropower-works (accessed 2023-10-24).
7. dhirajkumar612. 👨‍🎓Kaggle Survey 2022📄✍📊. https://www.kaggle.com/code/dhirajkumar612/kaggle-survey-2022/notebook (accessed 2023-10-24).
8. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **2001**, *29* (5). DOI:10.1214/aos/1013203451.
9. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Li, C.-L.; Kurakin, A.; Cubik, E. D. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems* **2020**, *33*, 596–608.
10. Kim, J.; Min, Y.; Kim, D.; Lee, G.; Seo, J.; Ryoo, K.; Kim, S. Conmatch: Semi-Supervised Learning with Confidence-Guided Consistency Regularization. *Lecture Notes in Computer Science* **2022**, 674–690. DOI:10.1007/978-3-031-20056-4_39.
11. Duan, Y.; Zhao, Z.; Qi, L.; Wang, L.; Zhou, L.; Shi, Y.; Gao, Y. Mutexmatch: Semi-Supervised Learning with Mutex-Based Consistency Regularization. *IEEE Transactions on Neural Networks and Learning Systems* **2022**, 1–15. DOI:10.1109/tnnls.2022.3228380.
12. Fan, Y.; Kukleva, A.; Dai, D.; Schiele, B. Revisiting Consistency Regularization for Semi-Supervised Learning. *International Journal of Computer Vision* **2022**, *131* (3), 626–643. DOI:10.1007/s11263-022-01723-4.
13. https://nsrdb.nrel.gov/data-sets/us-data (accessed 2023-10-24).
14. Hall, I. J. In Generation of a typical meteorological year; Sandia Laboratories, 1978.
15. Kalogirou, S. A. Environmental Characteristics. *Solar Energy Engineering* **2014**, 51–123. DOI:10.1016/b978-0-12-397270-5.00002-9.
16. Polo, J.; Alonso-Abella, M.; Martín-Chivelet, N.; Alonso-Montesinos, J.; López, G.; Marzo, A.; Nofuentes, G.; Vela-Barrionuevo, N. Typical Meteorological Year Methodologies Applied to Solar Spectral Irradiance for PV Applications. *Energy* **2020**, *190*, 116453. DOI:10.1016/j.energy.2019.116453.
17. Yang, H. Study of Typical Meteorological Years and Their Effect on Building Energy and Renewable Energy Simulations. *ASHRAE Transactions* **2004**, *110* (2), 424–431.

18. U.S. Climate Normals. https://www.ncei.noaa.gov/access/us-climate-normals/#dataset=normals-hourly&timeframe=30 (accessed 2023-10-24).

19. Yang, L.; Lam, J. C.; Liu, J. Analysis of Typical Meteorological Years in Different Climates of China. *Energy Conversion and Management* **2007**, *48* (2), 654–668. DOI:10.1016/j.enconman.2006.05.016.

20. Climate Change Indicators: U.S. and Global Temperature. https://www.epa.gov/climate-indicators/climate-change-indicators-us-and-global-temperature (accessed 2023-10-24).

21. Letcher, T. M.; Scott, K. In *Comprehensive renewable energy*; Elsevier, 2022.

22. Wind Energy Maps and Data. https://windexchange.energy.gov/maps-data (accessed 2023-10-24).

23. Solar Resource Maps and Data. https://www.nrel.gov/gis/solar-resource-maps.html (accessed 2023-10-24).

24. Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24* (2), 123–140. DOI:10.1007/bf00058655.

25. 1.10. Decision Trees. https://scikit-learn.org/stable/modules/tree.html#classification (accessed 2023-10-24).

26. 1.11. Ensemble Methods¶. https://scikit-learn.org/0.16/modules/ensemble.html (accessed 2023-10-24).

27. Arya, N. *KDnuggets*. 2022.

28. Sklearn.Ensemble.Randomforestclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed 2023-10-24).

29. Breiman, L. *Machine Learning* **1996**, *24* (1), 41–47. DOI:10.1023/a:1018094028462.

30. QuantDare. 2020.

31. Texas Land-Based Wind Speed at 100 Meters. https://windexchange.energy.gov/maps-data/357 (accessed 2023-10-24).

32. Global Horizontal Irradiance National Solar Radiation Data Base Physical Solar Model; 2018.

Authors

Phoenix Sheppard is a senior at the High School for Math Science and Engineering. He hopes to major in computer science and machine learning. He plans on pursuing a career as a software developer or machine learning researcher working to make the world a better place.

General guidelines:

1- Please read these instructions carefully and follow the suggested style.

2- If all contributors are from the same institution (high school) in that case no need to number each contributor.

3- Leave a space after corresponding author email and main text.

4- There is 8 pt after text spacing after each title. Please make sure to keep the template as it is.

5- Headings, such as Abstract, Introduction, etc, must be bold, Arial, 14 size with 8 pt after text spacing (Style 1).

6- The body text font size is 10 Arial and single spaced (Style Normal)

7- The following headings must be included: Abstract, Keywords, Introduction, Results and Discussion, Conclusion, Acknowledgements, References, Authors. Other subsections are optional and can be added by following the guidelines.

8- If you would like to add subheadings, their title should be italized and bold in 12 font size with 8 pts after text spacing (Style 2).

9- Figures and tables can be in single column or double column. Figures and tables should be located at the top or bottom of the column following their first citation in the text (unless they are equations, which appear in the flow of the text). **<u>All figures, images, and formulas need to be emailed as a separate attachment saved in .png format in high resolution.</u>**

10- Table caption should be placed on top of the table (see sample):

Table 1. Caption of the table (font 11, with 8 pts after text spacing), keep only first word capitalized, and end with a period.

P.S. Tables must be referred in the article text with their numbers (see examples in common mistakes section below)

11- Figure caption should be placed after the figure graph (see sample)

Figure 1. Caption of the figure (font 11, with 8 pts after text spacing), keep only first word capitalized, and end with a period.

P.S. Figures must be referred in the article text with their numbers (see examples in common mistakes section below)

12. When ready, research article should be named as "**LastName_FirstName**" or "**LastName_LastName_LastName**" (if multiple authors) and submitted as single pdf. Any images should be included in article as well as the originals should be attached to email separately in high resolution and .png format.

13. You may suggest reviewers for your paper. Please submit the following page with your paper.

Common Mistakes:

1. Include the School Address in addition to School Names on the article,

2. Include an email address of the corresponding author (one email) after the address on the article.

3. Keyword selection: The first two keywords MUST be from the category list. The first keyword should the general category and the second keyword should be subcategory, and other three should your own keywords. There should not be more than FIVE keywords.

4. All Tables and Figures MUST be referred in the text, at least like in parenthesis where the table figure data is discussed (Table XX).

5. Table and Figure names in texts must start with capitalized letters. Eg. As shown in Table 3, cats are better than dogs, but Figure 1 shows that dogs can be as cute as cats. There is other evidence (Table 4) which may suggest that fish in a tank can be a good home pet for kids.

6. Reference numbers in text MUST be placed after period or comma and MUST be superscript. Also, when continuous numbers are used as reference in the text, instead of putting all in, they should be places in using dash, such as 3-5 (instead of 3, 4, 5).

7. If you are not sure, please check our published articles as you prepare your manuscript for guidance.

Reviewer Suggestions

IJHSR requires authors to recommend 3 - 6 qualified reviewers in the same discipline the research is conducted in.

Please provide the first and last name, institution, city/state/country of institution, and email address for each suggested reviewer.

Recommended Reviewers must:

- Be fluent in English
- Be from different institutions (limited to 1 person per institution)
- Have not collaborated with the author's work
- Not be affiliated with the author's institution
- Should have PhD in the research area or more than ten years of experience in the field

First Name and Last Name	Institution	City/State/Country	Email
1.			
2.			
3.			
4.			