Predicting GHI using webcam images and meteo data

Jade Therras, Alison Bans, Antonin Mignot
ML4Science in collaboration with LAPD, CS-433, EPFL, Switzerland

I. Introduction

It is in society's best interest to transition to more sustainable energy sources and optimise their generation and distribution. In Switzerland, the electrical power generated by photovoltaic installations alone produced 3858 GWh in 2022 compared to 299 GWh in 2012 [4]. Therefore, the impact of solar energy on the Swiss power grid becomes more and more significant. As storing great amounts of electrical energy is impossible for now, energy companies are looking for new methods to predict the power generated/consumed in advance to tune their grid accordingly. Solar panel performance is inherently tied to solar radiation levels, referred to as Global Horizontal Irradiance (GHI). Having an accurate and localized prediction of the GHI for the coming hours could thus be very beneficial.

To this day, GHI predictions are predominantly supplied by meteorological companies that rely on satellite imagery [9] and complex algorithms. [1] However, these predictions often lack spatial resolution i.e. they are not precise for small areas. Moreover, they may not be highly accurate for specific timeframes. Consequently, the objective of this project is to develop a machine learning algorithm capable of generating accurate local predictions of the GHI two hours after the sampling of meteorological data (such as wind, current GHI, date, etc.) and webcam images sourced from two cameras situated on the EPFL campus. As the goal is to predict the GHI accurately, this is a regression problem.

This project was made in collaboration with the LAPD (Laboratory of Applied Photonics Devices) [5] in the context of the ML4Science project proposed by the Machine Learning course (CS-433) of EPFL. The lab provided a first model (in Keras) as a baseline model to improve, as well as a dataset and indices corresponding to three validation sets carefully chosen to be representative of the whole data. Three distinct model architectures were built on this basis. The first two models (A and B, III-A and III-B) use a combination of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) units and combine images and meteo data in distinct ways. The third one is based on an image-free dataset and therefore only relies on Multi-Layer Perceptron (MLP) combinations with LSTMs. Finally, an entirely different architecture has been put in place to make a fourth model (D, III-D) which uses vision transformers [6] on the images and finally combines it with the meteo data with MLP layers.

II. Dataset and visualisation

A. The Datasets

Two different datasets were provided by the LAPD to tackle our task. In the initial set, each datapoint consists of two RGB images, weather data (wind speed, wind direction, temperature, GHI value), and a time stamp (day, month, year, hour, minute) as represented on figure 1. The pictures and the GHI values are collected every 10 minutes from two webcams on the EPFL campus. On the other hand, the meteo data has been collected by Meteo Suisse in Lausanne. As they both come from different entities, it is important to align them, handle missing values and create the labels. Indeed, the cameras sometimes lag and therefore do not collect any images. The data points at these times must therefore be discarded. Similarly, points with missing meteo values are also removed from the set. It is also important to note that, due to the goal of this study and the file size of the images, all night-time measurements are discarded. The 360° wide images were then converted into 250x250 squares. Hence, a thorough initial processing was done before we received the data.
A second, more complete, dataset, was provided to us later as we wanted to create a time dependence between the data which was previously hard to do as the set was not entirely continuous. The alignment, handling of missing values and generation of the labels had to be done by us due to the raw nature of the files.

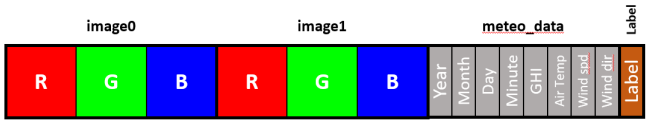


Figure 1: Structure of a datapoint

B. Exploratory data analysis

Visualisation has helped us understand different possible biases and imperfections of the dataset. An initial finding showed a total lack of data in April due to a fire that happened on campus15. Our model might therefore perform terribly for this month. Another important bias is that of the labels. As the data points start each day at sunrise but the labels representing GHI values start only two hours later, we are missing this 2 hour interval of non-null predictions. However, since the last image in a day is taken at sunset,

the two final hours of predictions are almost at zero (no sun)14. We could also observe that June, July and August share similar patterns in temperature, wind speed and wind direction A. Furthermore, they represent approximately 40% of the whole dataset. The model might thus overfit summer months. Some outliers can be found in the meteo data as well (wind direction, wind speed and temperature). However, we will not try to change them as we believe that they represent accurate values of unusual weather conditions. [7]
The visualisation of the raw GHI file of the second set showed many outliers (e.g. high GHI values at night, or any value going above 1000) which has allowed us to adjust them.
Finally, the dataset covers only one year. Therefore, it becomes difficult to extract data for validation while leaving some similar data for training, especially since 40% of the dataset is based on summer conditions.

C. Validation sets

After a discussion with the LAPD, it was told that a test set could be provided from new data. However, due to a miscommunication, said data was not provided thus no model could be tested. Therefore, the validation sets were used as test sets.
As the dataset is spread over a year, which means it is both influenced by seasonality and time dependence within each day, the validation sets were not chosen from random sampling as usual machine learning practices, but three validation sets were manually formed. Each set is made from consecutive timesteps and is not shuffled. The first validation set [27/01:10/02] represents the winter climate. The second [07/07:12/07] represents summer sunny days. Finally the third [18/09:25/09], is a more "challenging" set as it comprises cloudy days. The validation sets might not be exactly representative of the whole dataset thus achieving significantly different losses compared to the train set. Interestingly, the validation loss was lower than the training loss, which is not an unusual behaviour for neural networks [8].

D. Dataloaders

As these models are all time-dependent, the way the data is fed into the networks is very important. In the model received from the LAPD, the time dependence seemed lost as the samples were shuffled. The pictures were then fed into the network one at a time. While this works well, we still implemented a time window system to see whether it would improve the system.

Let us introduce the parameter L representing the "look-back" inspired by Tam's implementation [10]. At each step, the network is fed with a given time t, $t-1$, $t-2$, ..., $t-L$. 18 While this drastically increases the computation time, this method allows the LSTMs to "understand" the time dependence between each timestep. Each of those batches elements is randomly sampled during training, but not during validation, as it is more suitable for this context.

III. Models

A. Base model : with meteo data included after LSTM

As an initial model, we decided to re-implement the Keras code which was given to us by our supervisor in PyTorch and complement it with the meteo data. The model is made of two layers each applying 2 dimensional Convolutional Neural Networks (CNN) with a Rectified Linear Unit (ReLU) activation function to the images followed by batch normalisations. They are then fed into a 2-dimensional max pooling and some random dropouts are implemented. Two sequential Long Short-Term Memory (LSTM) are then applied to these webcam images. The output is then fed to a fully connected Multilayer Linear Perceptron (MLP) with a ReLU activation. Separately, the meteo data goes through two MLP layers with a ReLU activation function. In the final step, the outputs of both neural networks are combined and adjusted with a final MLP.A

Cross-validation was done by training on the training set and validating on the 3 validation sets. The images were reshaped to 100x100 for computational cost reduction purposes and the training data was shuffled as it showed better results than keeping it in chronological order in previous exploration. A range of dropout value combinations and batch sizes were tested in jointly with the normalisation of the meteo data and by shuffling or not the validation sets. The results of this cross-validation showed very little variations in the validation loss but we were able to notice that adding some small non-null dropout rates improved the model. The choice of whether the data points were shuffled or not in the validation sets did not show a significant difference although it makes more sense to keep it in the right order as the predictions would be made chronologically. We then further decided to try adding padding as this would not discard some potentially crucial information on the border of the images such as clouds entering the picture. Surprisingly, this did not show significant improvement and neither did the adjustment of outliers. However, we could observe a true improvement in the loss of the last validation set when dividing the labels by 100 to train the model.

In parallel, a very similar architecture was implemented with a custom feature: a cloud detector. Clouds are really important for GHI prediction, therefore the model must detect them efficiently. We implemented several cloud detectors based either on the level of colours (grey, white and optionally yellow or grey and blue) in the images or on edge detection, in order to help the model learn. Resulting binary features are added as a new dimension to each image (R, G, B, cloud), and the model is changed accordingly.A The cloud detection is not extremely accurate and is sensitive to sunlight, fog and artefacts, as it can be seen on figure 2.
Additionally, increasing the resolution of the images leads

to a longer training time and does not seem to improve the results. The predictions are also less stable, showing a lot of tremors. Adding the cloud filter to the RGB image does not show significantly better results either. Interestingly, the results are not much worse while giving only the cloud dimension to the network (no RGB). Although, the variance of the loss between each training renders it difficult to draw conclusions on small changes. A possible interpretation would be that the model already learns the cloud positions from the original images, making the detector unnecessary, but the notion of cloud is crucial to predict the GHI, making the cloud dimension sufficient to obtain an acceptable results.

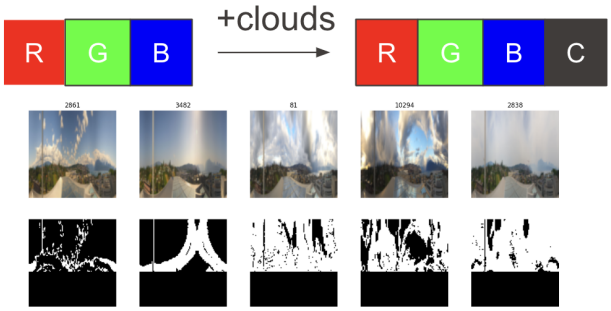


Figure 2: addition of the custom ”cloud” feature

B. Model with meteo data included before LSTM

The architecture of our second model strongly resembles the first one (A) with the slight change that the features extracted from meteo data are introduced into both LSTM layers with the image features A. Combining this with the rolling window dataloader should show better results than our initial model. Since we did not see any drastic changes when varying the dropout rates or by adding padding, we decided to focus more on the lookback (L) parameter. We expect that increasing it would allow the model to learn patterns in time. However, due to the drastically increasing computation time with bigger values for L, we were compelled to try with values of up to 3 only. We have also divided the labels by 100 which has shown great improvement in model A for the last validation set (September). The best results were obtained for $L = 2$ and with the division of the labels by 100. However, this model did not obtain the lowest loss on all three validation sets.

C. Model with only meteo/GHI

In this model, the second dataset has been used for its consistent time interval with which time dependence between consecutive points can be harvested. This time, we received the raw Excel files (meteo and GHI) of the data which needed some reshaping and adjusting.
Keeping only important meteo information, both data frames were merged by aligning their values according to the corresponding time stamps. An extra feature has been added by transforming the month to a cyclic value using the cosinus function. Finally, the labels were generated by shifting the GHI values of two hours. Validation setsII-C are extracted by following the timestamps of the sets given to us by the LAPD. More data points are present in those validation sets. The mean and standard deviation of meteo data from the training set were computed and exported to allow standardisation. Furthermore, the median and quantiles of each hour and minute of the year were computed for the labels to be able to adjust the outliers.
The network itself is relatively simplistic as it consists of two MLPs with a ReLU activation function followed by two LSTMs and 2 MLPs (one of which has a ReLU activation function) with a final dropout layer. A. For this model, the lookback window brings a significant improvement, and as no images are used, the computation time is still reasonable. Values of L going from 1 to 4 have been tested with 4 showing the best results.

D. Visual transformer model

We implemented a Visual Transformer (ViT) based on the work of Phil Wang from MIT [12]. Even though the code is used for classification by the source, we will be using it for regression and adjusting the 'number of classes' parameter as needed. We first applied a simple ViT to only one image and tried to predict the GHI. We subsequently implemented a model which takes the images from both cameras and processes them simultaneously through separate ViT. The resulting features are concatenated and passed through two MLP layers. The meteo data could then be added before the first or the second MLP, or not at all. A.

We also tested another model using SimpleViT, a variant of ViT which is claimed to be more efficient according to the source. Note that a lot of different ViT transformers could have been used, but due to time restrictions, we were unable to explore them more (some examples are included in the code). We could also have applied the same transformer to the two images concatenated similarly to what was done for the CNNs. [11]

E. Postprocessing

We implemented some post-network processing to improve the results of the network. First, we know that at night, the GHI prediction should be 0. Therefore, we made a manual processing to set those values to 0. This doesn't improve the quality of the prediction even if the error decreases, but it helps in comparing models.

Secondly, we observed a lot of noise in most of the model results. For the GHI prediction, the most important is the direction of the evolution (is the GHI increasing/decreasing and how much). We therefore tried to fit a polynom approximation to the predicted result using the Savitzky Golay filter. [3]

IV. Results and Discussion

Val. set	Aa	Ab	B	C	D
1	57.22	83.44	76.39	36.35	96.52
2	46.00	64.53	52.49	23.18	47.49
3	94.86	120.82	94.13	65.32	103.77

Table I: Summary of the best results for all the models

The table above presents the best results obtained with each model described previously. The ViT (D) seems to perform somewhat decently for the second validation set but terribly for the two others. As this was the last model implemented, we believe that it could be improved with further testing but it should already present a good basis.

Surprisingly, the model where a cloud filter is added to the RGB images (Ab) does not perform the best either. Its losses are much larger than its cloudless counterpart (Aa). As these two models were developed in parallel, they may not have the same basis, so putting them together might improve the results. For example, model Aa has different dropout rates and includes padding, but it also divides the labels by 100 before training which seems to be a crucial step in reducing the loss of the third validation set.

This transformation of the labels has also been done for the network in which the meteo data is fed into the LSTM layers (B) and it shows once more a lower loss for the set 3. However, it does not seem to improve the predictions for the two other sets (1&2) when compared to keeping labels to their original values (not shown here). Although this might seem like a limitation, we believe that it might be a strength of those models. Indeed, the dataset is a little biased towards good weather as June, July and August represent 40% of the whole data. And being able to decrease the loss for cloudy days seems to be the most important aspect of a network as they are the hardest to predict.

The network using only meteo data (C) seems to show the best results according to this table. According to the plots of figure 20, predictions for sunny days are very good. However, the validation sets contain more data due to their continuity and therefore also predicts GHI values throughout the night which is very easy. Hence, the excellent performance of the model in those time ranges induces a lower mean error making these numbers not comparable to the others.

On a more graphical analysis of the predictions (see figure 3 or 19), we can see that including the meteo data before the LSTM (B) leads to a small improvement of the network behaviour as the GHI predictions at night is more accurate than in the model with meteo data being included after LSTM layers.

Focusing more on the cloudy days of our seemingly best model (C), as can be observed in figure 4, shows that predictions do not seem to follow the trend of the labels.

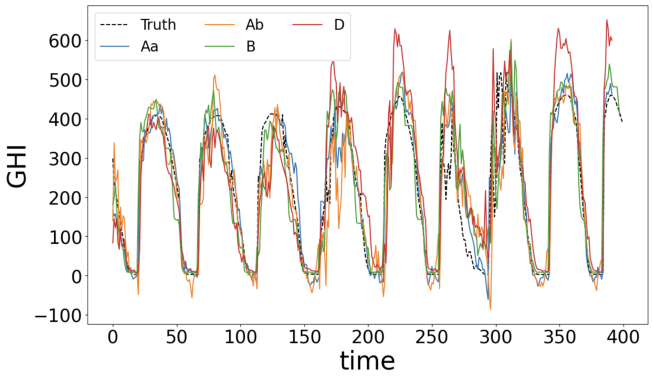


Figure 3: Superposed results for validation set 3

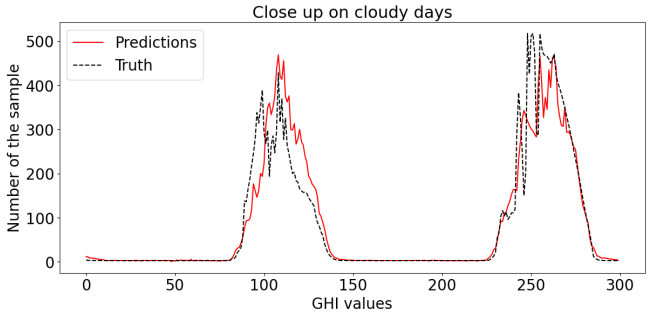


Figure 4: Two days with clouds that have bad predictions

V. Conclusion

The GHI prediction problem is tackled by many in the scientific community with the rise of the use of solar panels. However, localised predictions are still a challenge especially when weather is uncertain [1]. Our models may not perform the best predictions but small improvements have been observed in the multiple different tests and models. We are unable to name one best model as the one using only meteo data does not have comparable validation sets. However, since it learns time dependencies it would not be surprising that its predictions are the best. Ideally, we would like to see that the predictions follow the tendency of the labels even if they are not exactly correct. However, in our models, there are still important discrepancies.

There are different ways to obtain better results such as adding images from other cameras so the network can detect incoming clouds, or adding other relevant features. With another architecture, the ViT model could perform better. Dividing the data into days for training could also be a good idea as this would allow the shuffling between the different subsets (days). However, this would have been very challenging for our first dataset as the days vary in the number of data points they contain 16. Finally, collecting more data to feed to the networks could also be beneficial as the current set only spans one year.

References

[1] *All you need to know about GHI forecasting for solar PV plants*. en-US. Section: Article. URL: https://smarthelio.com/predict-solar-irradiance-with-ghi-forecasting/ (visited on 12/20/2023).

[2] *ChatGPT*. en-US. URL: https://chat.openai.com (visited on 12/19/2023).

[3] *Cookbook/SavitzkyGolay - SciPy wiki dump*. URL: https://scipy.github.io/old-wiki/pages/Cookbook/SavitzkyGolay (visited on 12/20/2023).

[4] Office fédéral de l'énerge. *Statistiques de l'énergie solaire*. French. Tech. rep. Office fédéral de l'énergie OFEN, Aug. 2023. URL: https://www.swissolar.ch/03_angebot/news-und-medien/statistik-sonnenenergie/statistique_energie_solaire_2022_rapport_fr_final.pdf (visited on 12/16/2023).

[5] *LAPD*. en-GB. URL: https://www.epfl.ch/labs/lapd/ (visited on 12/19/2023).

[6] *Papers with Code - Vision Transformer Explained*. en. URL: https://paperswithcode.com/method/vision-transformer (visited on 12/18/2023).

[7] *Réseau de mesures automatiques - MétéoSuisse*. fr. URL: https://www.meteosuisse.admin.ch/meteo/systemes-de-mesure/stations-au-sol/reseau-de-mesures-automatiques.html (visited on 12/20/2023).

[8] Adrian Rosebrock. *Why is my validation loss lower than my training loss?* en-US. Oct. 2019. URL: https://pyimagesearch.com/2019/10/14/why-is-my-validation-loss-lower-than-my-training-loss/ (visited on 11/29/2023).

[9] *Solcast — Solar Api and Solar Weather Forecasting Tool*. en. URL: https://solcast.com (visited on 12/20/2023).

[10] Adrian Tam. *LSTM for Time Series Prediction in PyTorch*. en-US. Mar. 2023. URL: https://machinelearningmastery.com/lstm-for-time-series-prediction-in-pytorch/ (visited on 12/20/2023).

[11] *Using ViTForClassification for regression? - Beginners*. en. Section: Beginners. Oct. 2021. URL: https://discuss.huggingface.co/t/using-vitforclassification-for-regression/10716 (visited on 12/20/2023).

[12] Phil Wang. *lucidrains/vit-pytorch*. original-date: 2020-10-03T22:47:24Z. Dec. 2023. URL: https://github.com/lucidrains/vit-pytorch (visited on 12/19/2023).

Appendix

I] Ethical risks

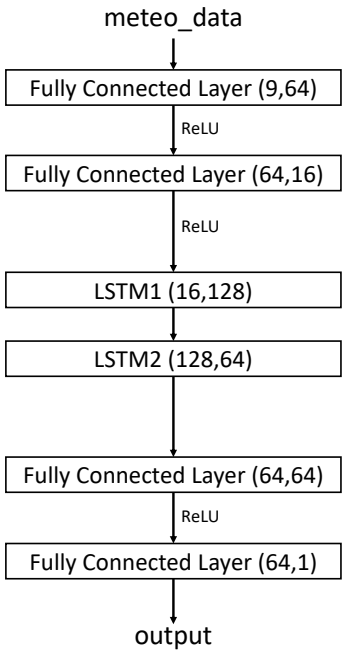
Since the output consists of GHI values (numbers), we do not think that they or their use would create some ethical issues. The biggest stakeholders that would benefit from our project would be the energy companies that manage the power grid. As far as fairness is concerned, those models could only be implemented in regions where meteo data is accurate and where public webcams are available, but also regions that notably use solar panels as an energy source. Therefore, it would benefit mostly rich regions but it is not the major issue.

With all that in mind, we identified an ethical risk in our project regarding the use of webcam images. Those images capture people wandering around the campus. The use of images from a public space is subject to regulations because people are not aware of the webcams and of their use. The main risk is for the security of the data. If a malicious person were to hack into our computers or our online storage, one would be able to recover images and exploit them.

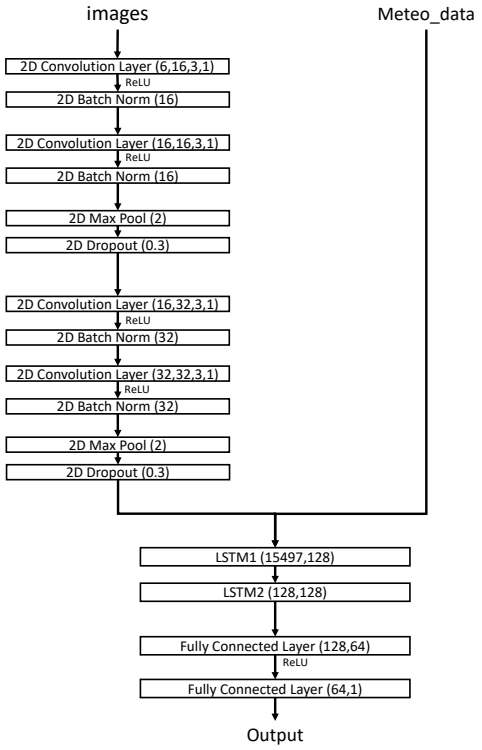
To tackle this problem, the images were previously cropped to keep just the sky. And we have also attempted to use only the clouds with the help of a cloud detector. In both cases, the results were unfortunately unsatisfying as the lower part of the image is still important to predict the GHI. To counter this, the images were resized instead to lower the resolution. People in the pictures become thus unrecognisable.

II] Model schematics

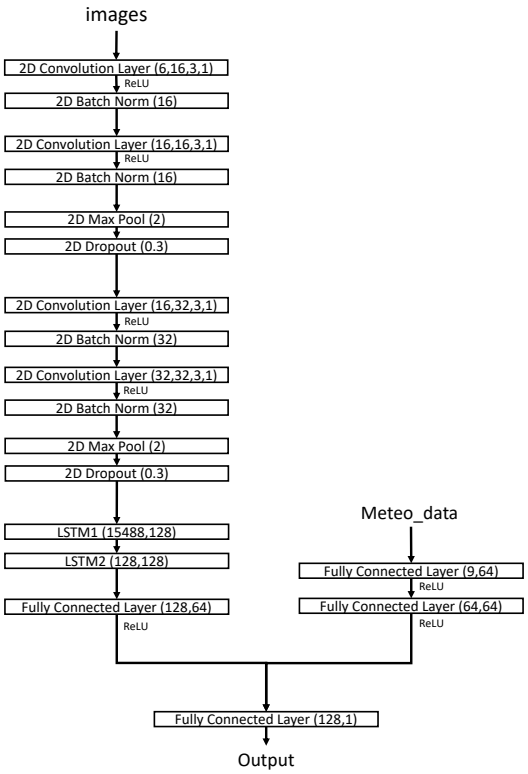
Model without images



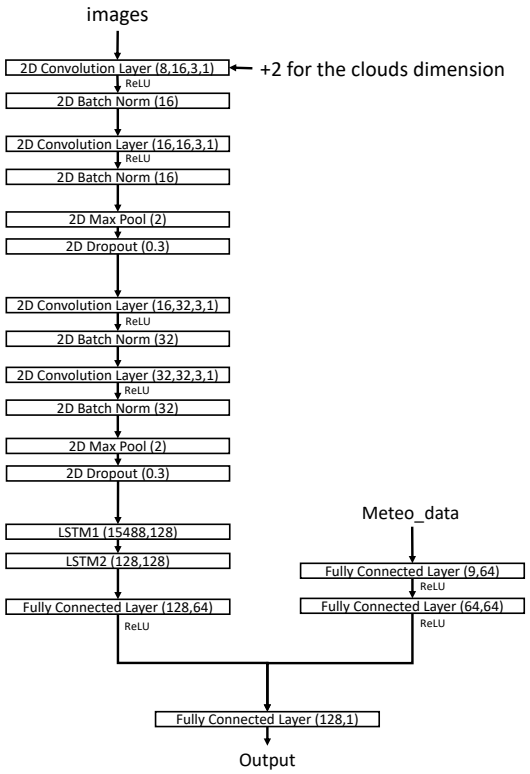
Model with meteo_data before LSTMs



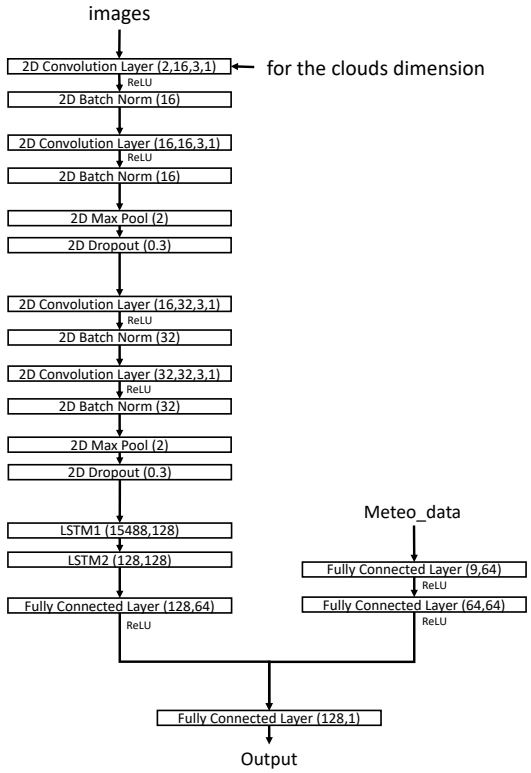
Model with meteo_data after LSTMs



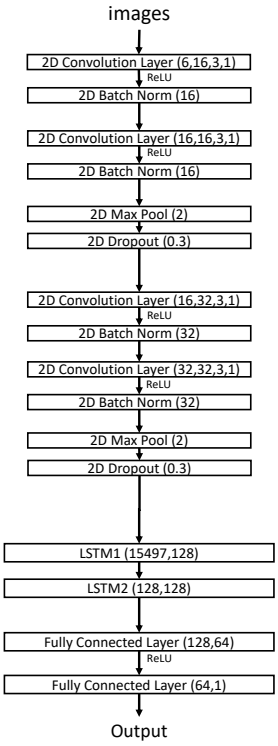
Model with clouds



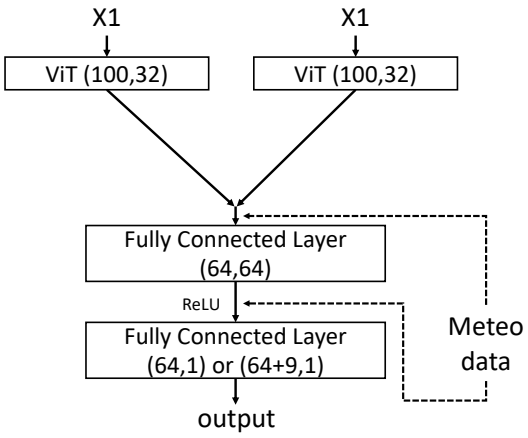
Model with only clouds



Model with only images



Vision transformer



III] Digital ethics canvas

Digital Ethics Canvas 2023 - Software Design version - C. Hardebolle, V. Macko, V. Ramachandran, A. Holzer, P. Jermann

Context	Beneficence	Non-maleficence		Solution
□ In which context is the solution evaluated? For optimal use of solar panels, it is important to predict accurately, locally and for short-term the "amount of sun" received, quantified using the GHI (Global horizontal irradiance). In function of the predicted level in 2 hours, the energy company will optimize its own to redistribute the production. This project will be used by energy companies. For now, the GHI in predicted using satellite data but meteo compagny. It is costely and not extremely accurate locally.	□ What are the expected benefits of the solution in this context? Low cost, local, easy to use, easy to implement (only need public meteo data and few webcam). Useful in the context of the energy transition.	**Risks**	**Mitigation**	□ What are the characteristics of the solution under evaluation? The prediction use local meteo data and webcam from the campus of EPFL and aims to predict the GHI in 2 hours locally better than the meteo.
		□ Can the solution be used in harmful ways, in particular with regards to vulnerable populations? Not really, only need to install webcams, but a solution for region with solar panels so not everyone. □ What kind of impacts can errors from the solution have? Inefficient use of energy, more loss, more cost □ What type of protections does the solution have against attacks or misuse? No real protections Need strong clear data protection, only a few people have acces to images before reduction/distortion.		
	Privacy		**Fairness**	
	Risks	**Mitigation**	**Risks**	**Mitigation**
	□ What data does the solution collect? GHI data □ Is it collecting personal or sensitive data? Picture from a public space □ Who has access to the data? The coder and developper have acces, we only have acces to reduced images so not identifiable people.		□ How accessible is the solution? Accessible as only webcams is needed but can only be implemented for solar panels. □ What kinds of biases may affect the results? Only test on one locality □ Can the outcomes of the solution be different for different users or groups? Not, maybe in different region but will be	
	□ How is the data protected? □ Could the solution disclose / be used to disclose private information? When people are identifiable, take the images away (will need a complexe processing)		retrained anyways. Maybe less performant in cloudy and extreme/rapid change environnement. □ Could the solution contribute to discrimination against people or groups? Not, depend of the use. May reduce energy cost in region that have access ?	
	Sustainability		**Empowerment**	
	Risks	**Mitigation**	**Risks**	**Mitigation**
	□ What is the carbon footprint of the solution? Carbon footprint of webcams ? but all the data is already use of other purposes. □ What types of resources does it consume (e.g. water) and produce (e.g. waste)? Energy for computers, processing, webcam. □ What type of human labor is involved? No big human labor to produce the data		□ Can users understand how the solution works and what its limits are? Yes, but have to precise □ Are users able to make choices (e.g. consent, settings) in their use of the solution and how? No need of that ? □ How does the solution affect user autonomy and agency? Helps to choose how to manage energy.	

Digital Ethics Canvas 2023 - Data Science Version - C. Hardebolle

Dataset	Beneficence	Non-maleficence	
		Risks	**Mitigation**
□ Who created the dataset? The local laboratory, from webcam of epfl and MeteoSuisse. □ For what purpose was the dataset created? For this project □ What mechanisms or procedures	□ What are the expected benefits of analyzing this data? For whom? Have good GHI prediction with a simple dataset. For solar panel company, energy company, for epfl (use of solar panel on the campus)	□ Does the dataset contain unsafe data (violence, nudity…)? No. □ What kind of impacts can errors in the data or in the analysis have? Bad prediction	

<table>
<tr>
<td rowspan="6">were used to collect the data?
Collect images every 10 min of the campus and match to meteo data.
□ Who was involved in the data collection process?
The laboratory (and meteosuisse)
□ Over what timeframe was the data collected?
Approximately 1 years, from january 2022 to september 2023
□ Was any preprocessing of the data done?
Only size reduction of the images. The GHI is put to 0 at night.
□ Are there any missing data or data errors?
One month is missing (april) because of a problem in the campus. Some other images are missing along the dataset (we use only the timeframe where we have all the data). We have a few outliers at night.
□ Where is the data stored?
On google drive.</td>
<td colspan="2"></td>
<td colspan="2">□ Could the data or the conclusions from the analysis be used in harmful ways?
Not really
Careful about the use of the data</td>
</tr>
<tr>
<td colspan="2">Privacy</td>
<td colspan="2">Fairness</td>
</tr>
<tr>
<td>Risks</td>
<td>Mitigation</td>
<td>Risks</td>
<td>Mitigation</td>
</tr>
<tr>
<td>□ Does the data contain personal or sensitive information?
The data contains picture of people (students, professor, collaborator, visitor ect.. in a public space)
□ Can personal or sensitive information be derived or inferred from the data or from the analysis?
Yes, normally people aren’t take in account in the analysis and the picture is too deformed/far for identification, but still we don’t have personnel consent and people may be identificable before data processing (we personally only have acces to the deformed images)

data protection</td>
<td></td>
<td>□ Is the data representative from a larger set (population)? How are subgroups represented?
EPFL population in the pictures, and data only from epfl so the model can only predict locally.
□ What kinds of biases may affect the data?
Biases of buldings and local meteo from the pictures ?
□ Can the outcomes of the analysis be different for different groups?
Not really
□ Could the data or analysis results contribute to discrimination against people or groups?
If the data is use for other purpose maybe, but for this purpose not.</td>
<td></td>
</tr>
<tr>
<td colspan="2">Sustainability</td>
<td colspan="2">Empowerment</td>
</tr>
<tr>
<td>Risks</td>
<td>Mitigation</td>
<td>Risks</td>
<td>Mitigation</td>
</tr>
<tr>
<td>□ What is the carbon and water footprint generated by the storage of the data and by the computation in the analysis process?
Small model and relatively small dataset -> not that much</td>
<td></td>
<td>□ How are the people concerned involved with the data or the analysis: have they been notified, have they consented?
EPFL as been notified but no consent for the people. It is a public space and people aren’t</td>
<td></td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>□ What type of human manual labor is involved in the data (e.g. labeling)?
No manual labeling or much processing. Will need more processing if we need to take off peoples
□ Does the data or the analysis require updates?
Not</td>
<td>identificable
□ Are the people concerned able to make choices (e.g. revoke consent, modify or delete data) regarding the data or the analysis?
EPFL have this power but not individual peoples.

Work for consent with EPFL, prefer data with no human in the pictures.</td>
</tr>
</table>

Risk Mitigation

IV] Data Visualisation

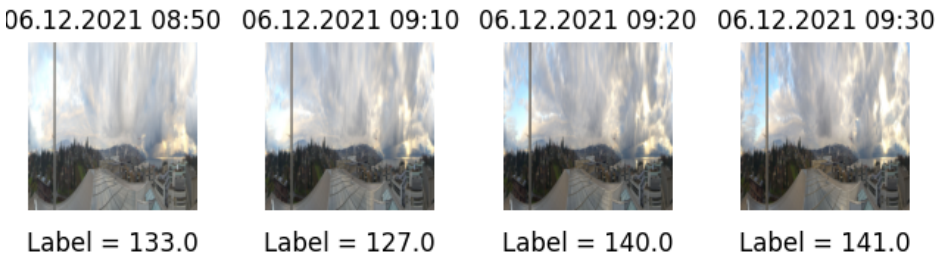


Figure 5: This figure represents four pictures captured by one of the cameras at different times. The labels represent the GHI values 2 hours after the respective image is taken.

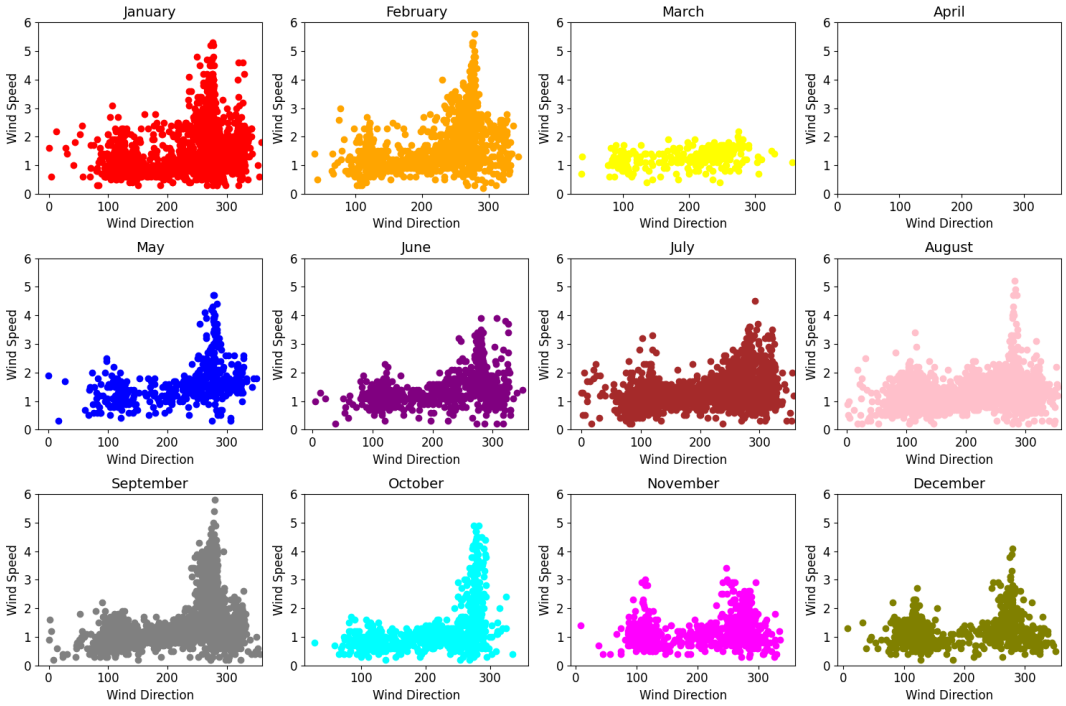


Figure 6: In this figure, a scatter plot of the wind speed vs wind direction has been done for each month individually.

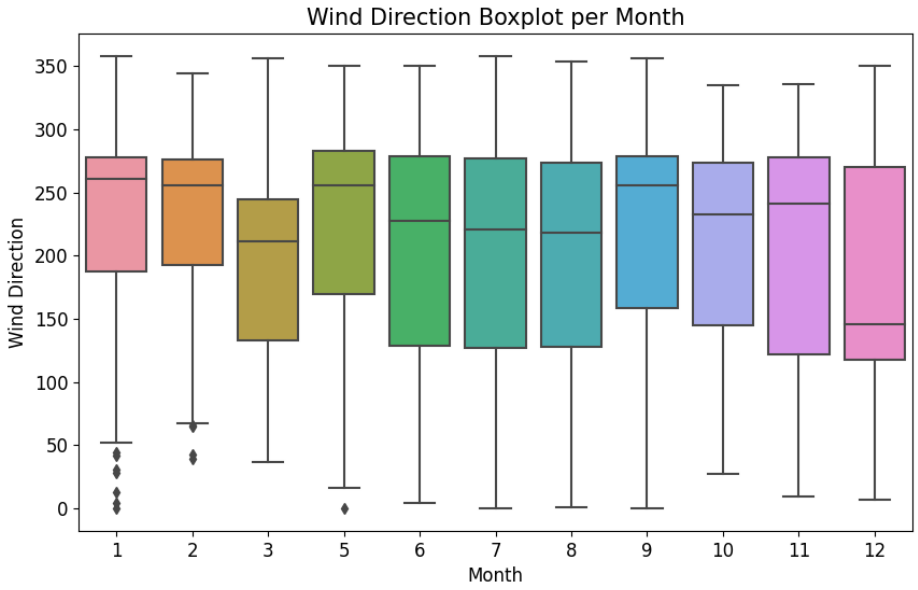


Figure 7: This image shows a boxplot of the wind direction for each month. Some outliers can be observed but due to the very reliable source that is MeteoSuisse, they were not adjusted in our models.

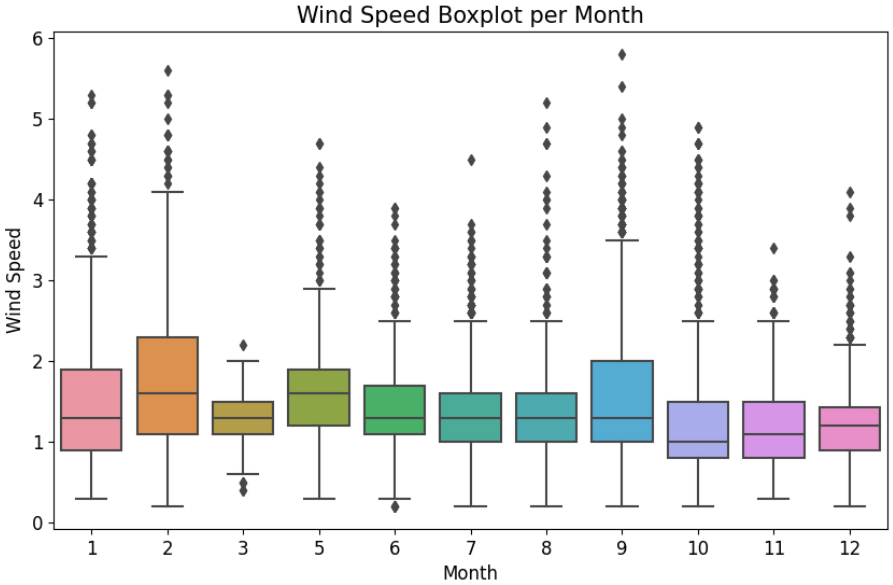


Figure 8: This image shows a boxplot of the wind direction for each month. A certain number of outliers can be observed but due to the very reliable source that is MeteoSuisse, they were not adjusted in our models. They can be explained by the presence of extrem weather conditions. We could indeed observe that those outliers usually occur on the same day and are consecutive.

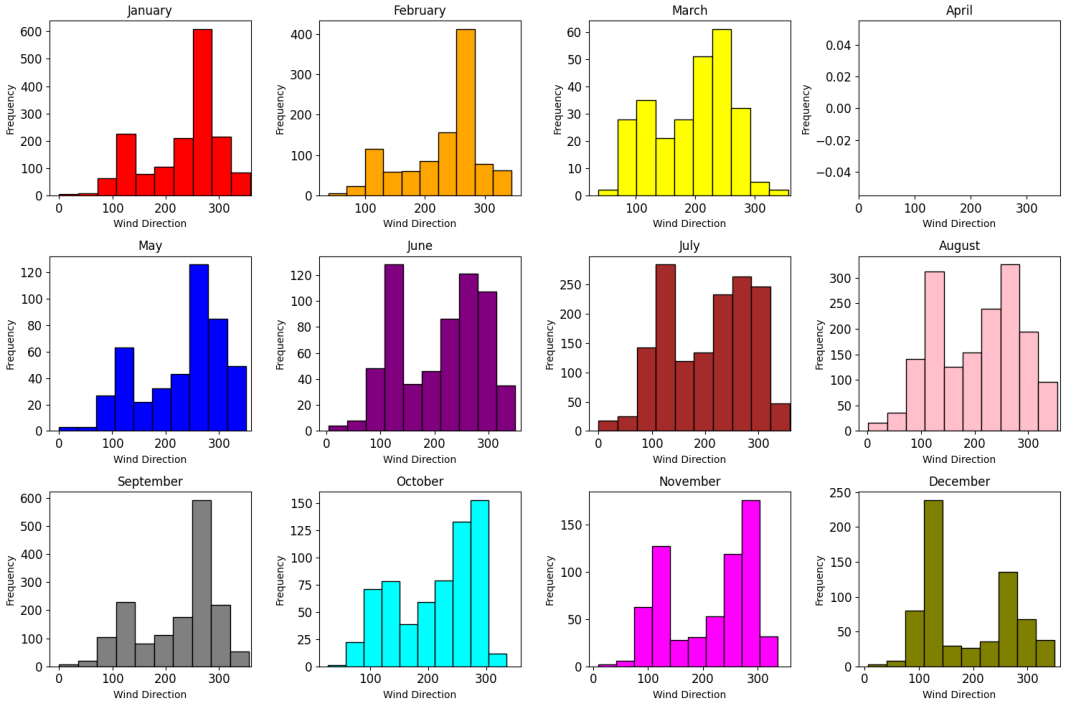


Figure 9: This plot shows the distribution of the wind direction for each month.

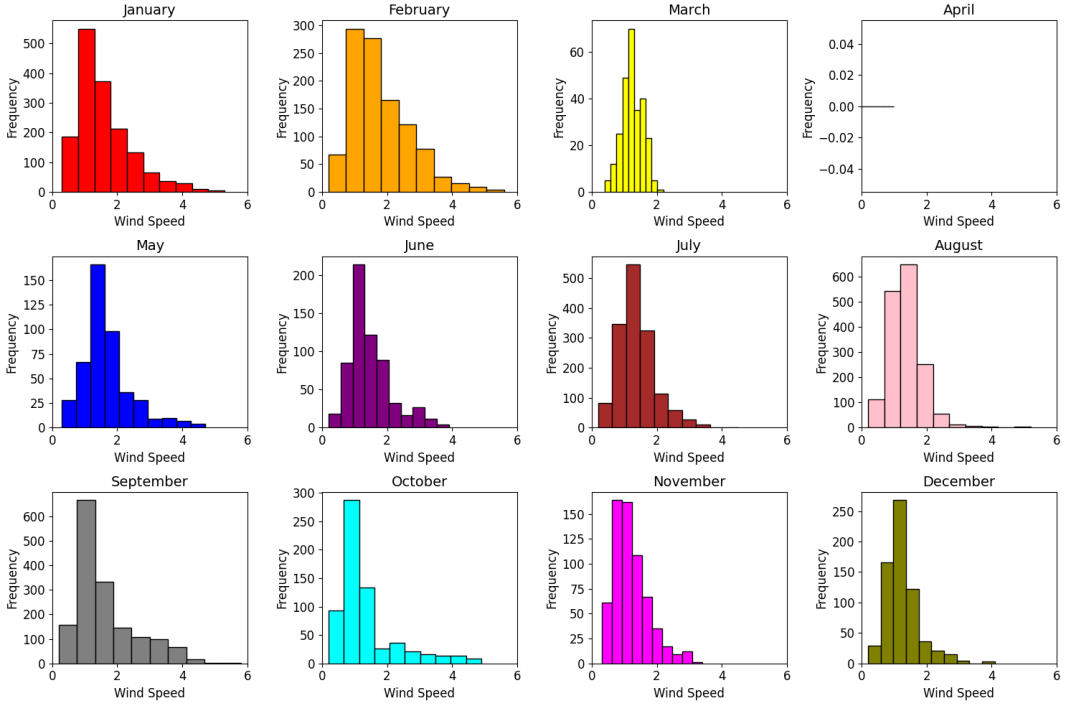


Figure 10: This plot shows the distribution of the wind speed for each month.

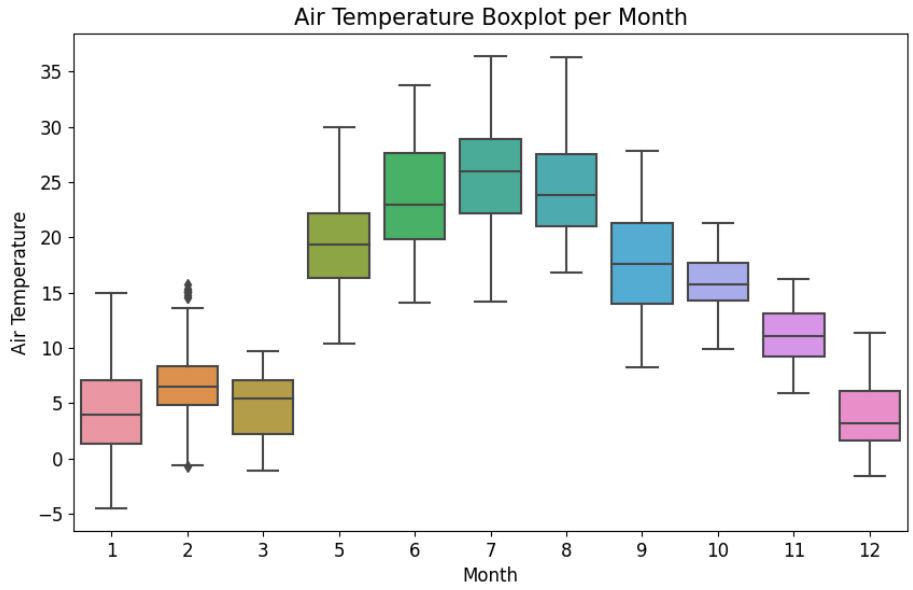


Figure 11: This figure shows a boxplot of the temperature for each month. A discontinuity can be observed between March and June due to the lack of data in April.

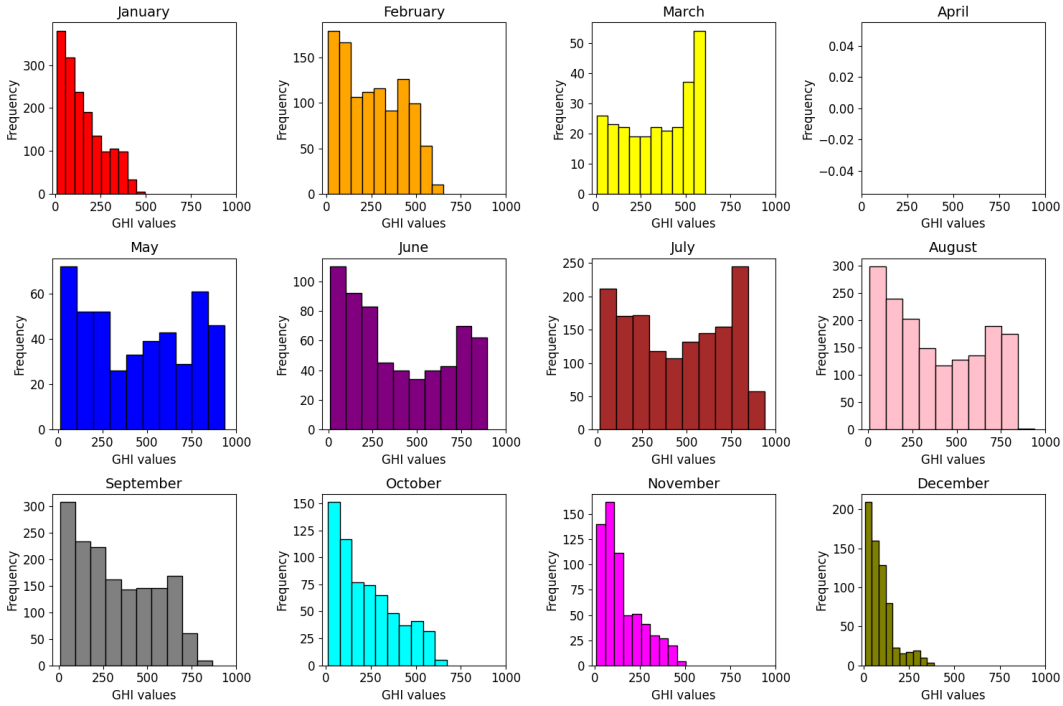


Figure 12: This image shows the distribution of the GHI for each month according to ground truth values not labels.

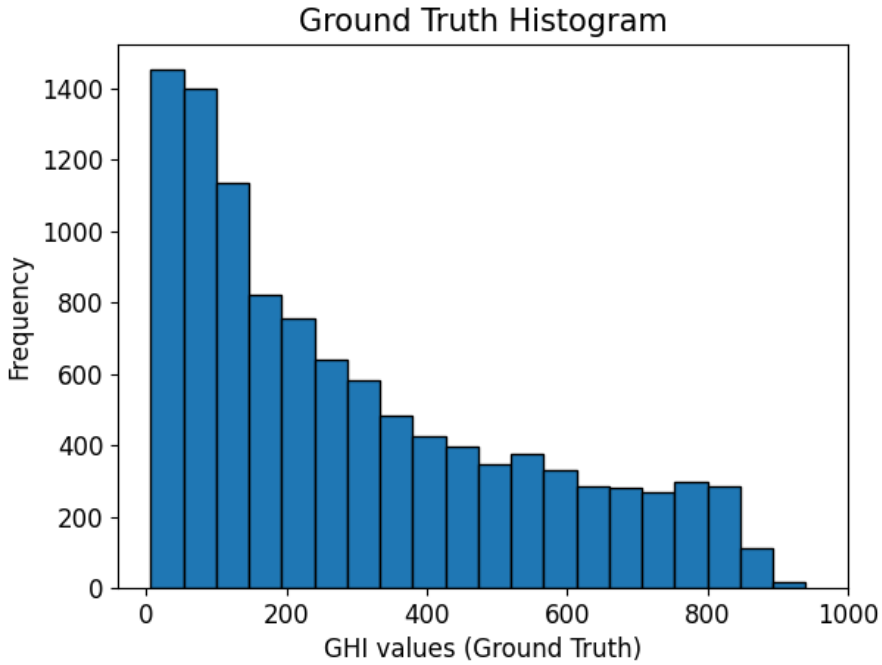


Figure 13: This image shows the distribution of the GHI for each month according to ground truth values not labels.

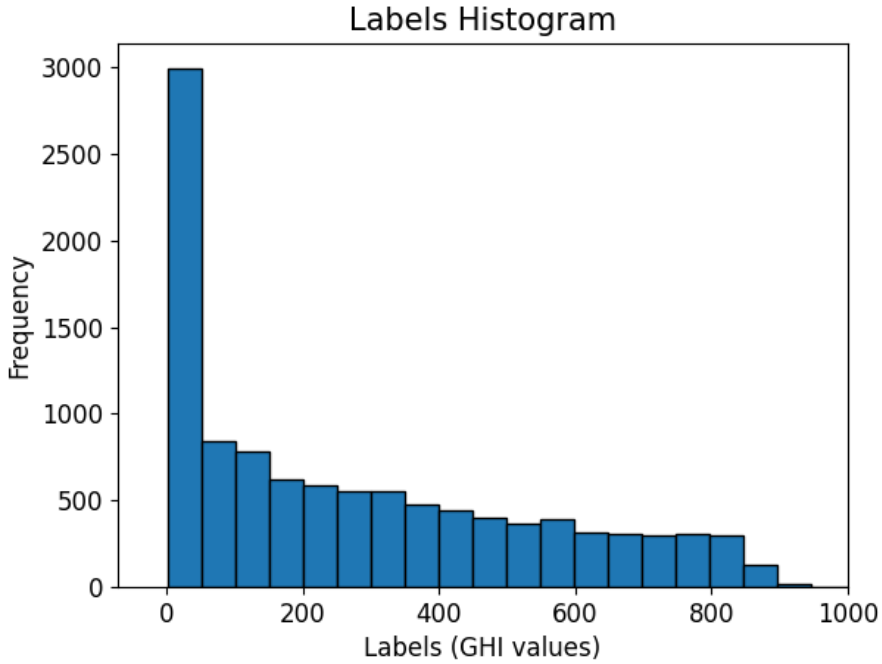


Figure 14: This image shows the distribution of the GHI for each month according labels (two hours after the data point).

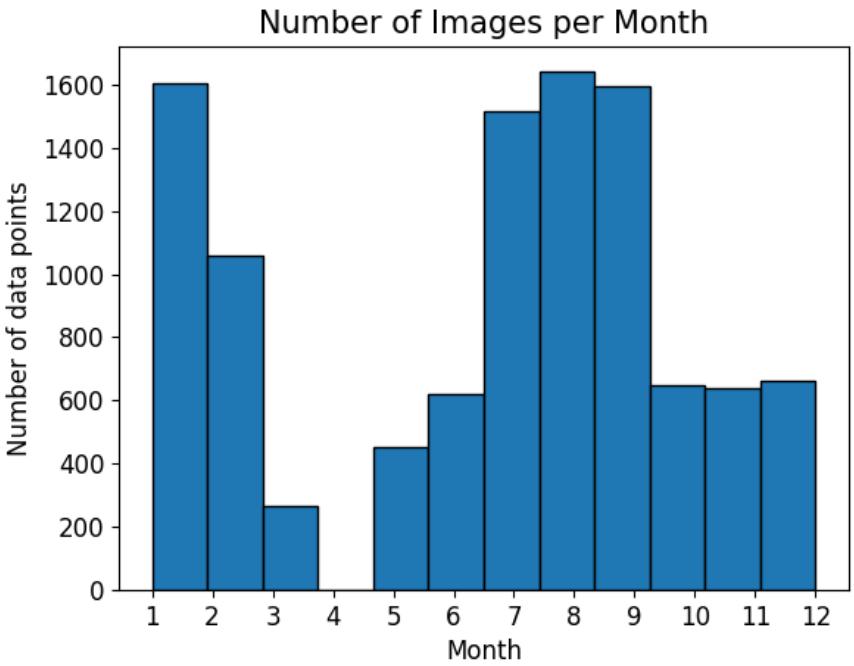


Figure 15: This image shows the the number of data per month.

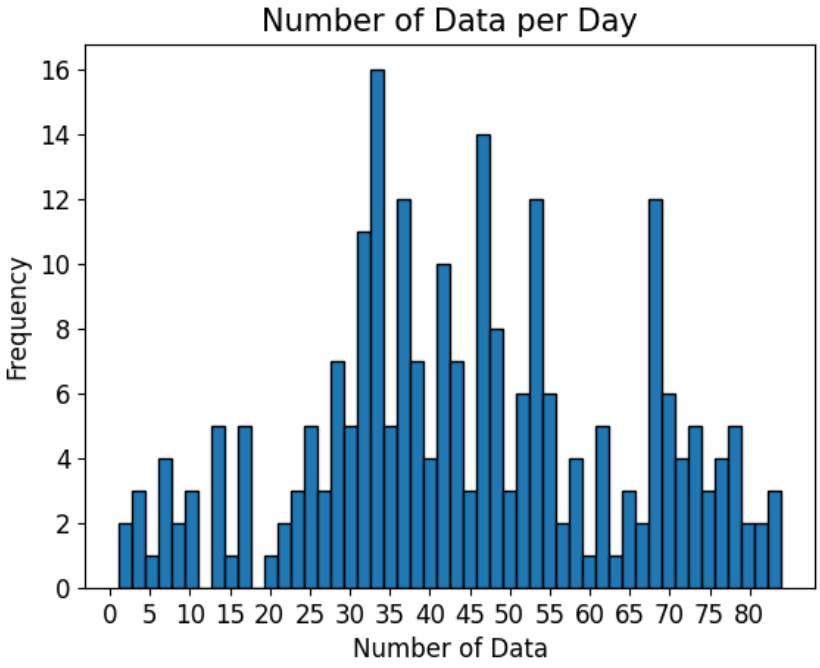


Figure 16: This figure shows the histogram of the number of data collected in a day.

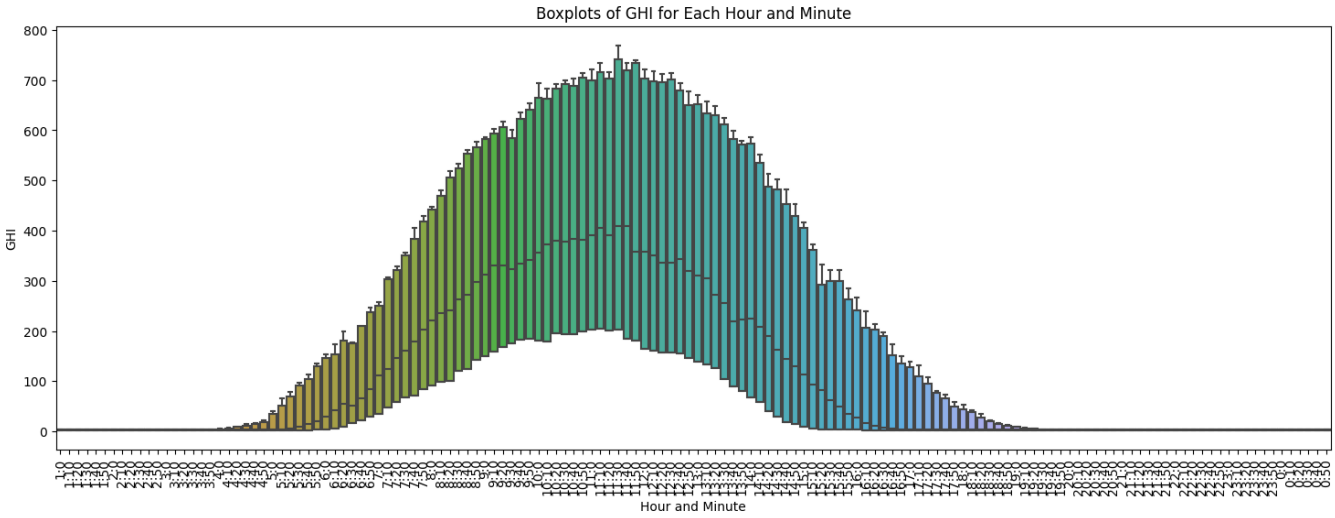


Figure 17: This image shows the boxplots of the GHI values for time of the day. It is based on the second data set received after adjusting the outliers.

V] Data loader

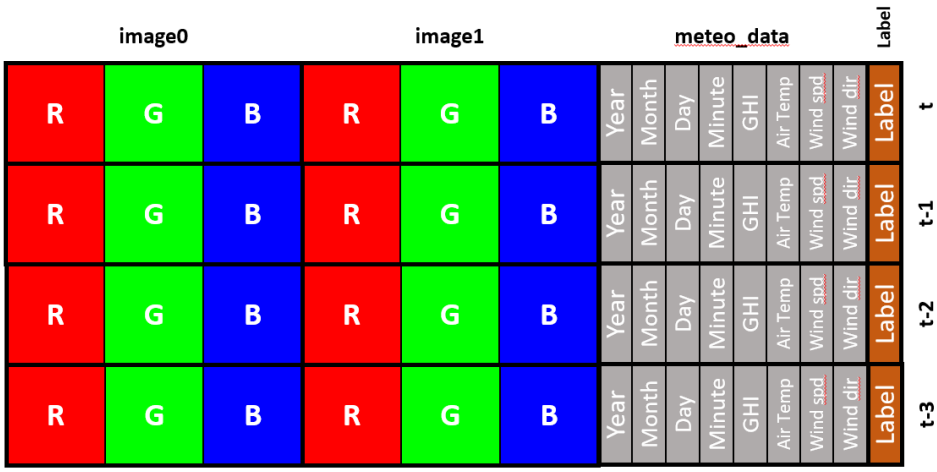


Figure 18: dataloading for L = 3

VI] Result

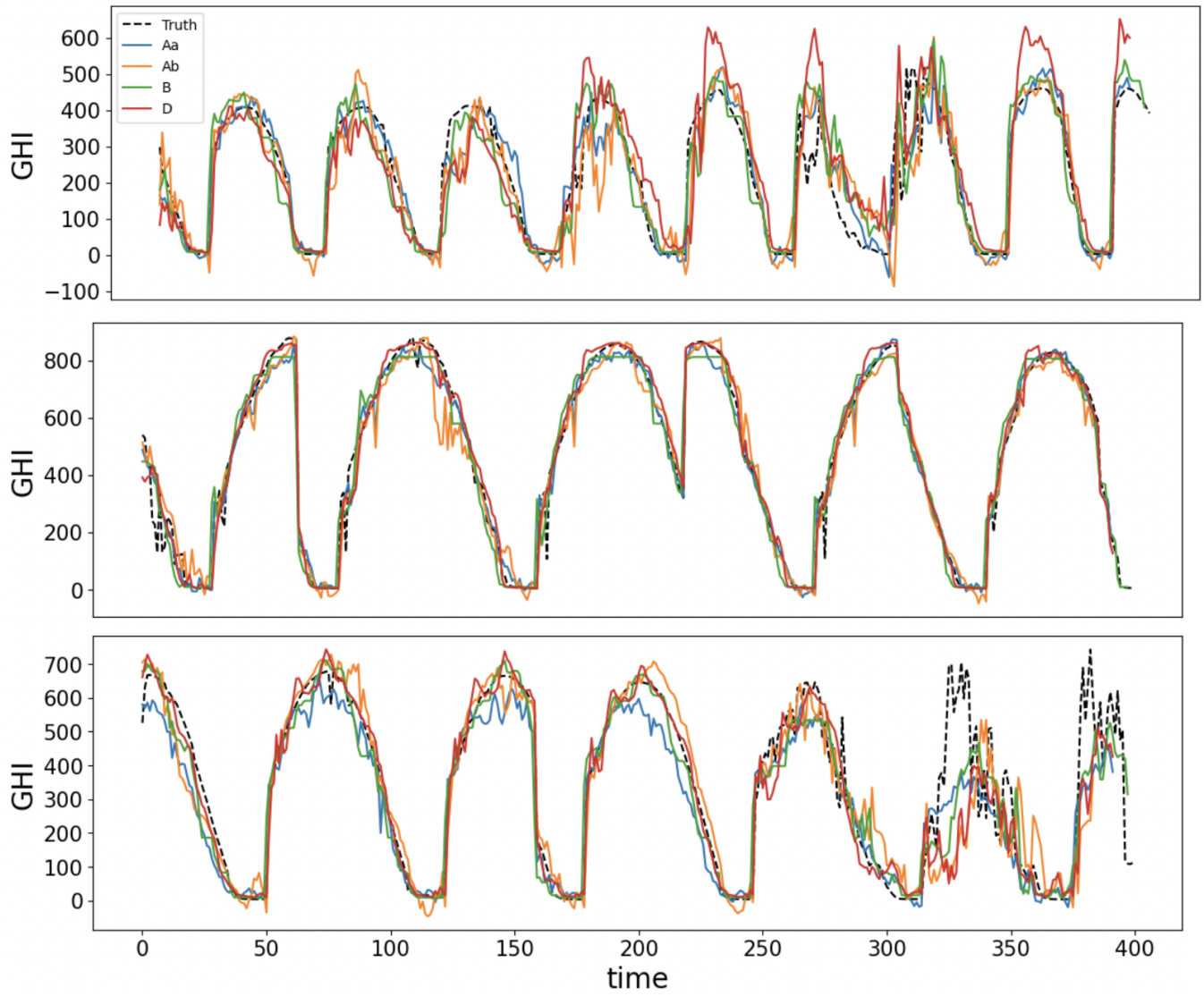
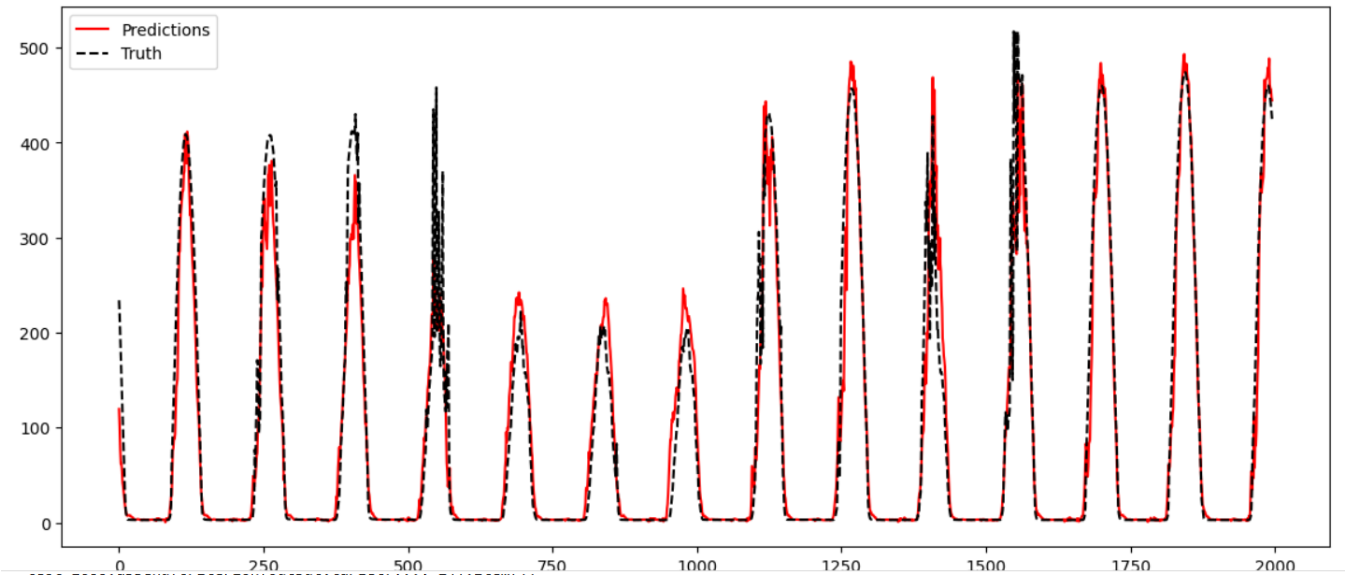


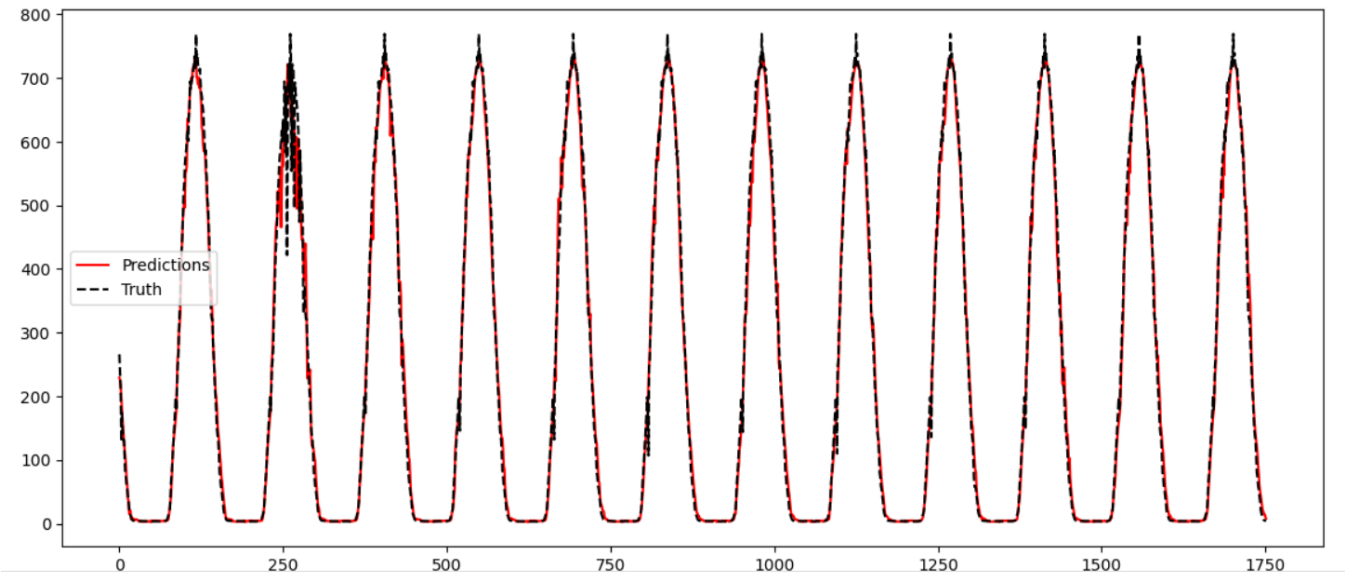
Figure 19: result for all the models

VI] Result

```
the test loss is  36.34922883765391
<matplotlib.legend.Legend at 0x7decf2bb9ff0>
```




```
the test loss is  23.179623148983172
<matplotlib.legend.Legend at 0x7decf2f09f00>
```




```
the test loss is  65.32309077044069
<matplotlib.legend.Legend at 0x7f9f2c005990>
```