Hourly Global Solar Radiation Reconstruction Applying Machine Learning

Francesco Mercaldo[*†], Antonella Santone[‡], Francesco Tariello[*], Giuseppe Peter Vanoli[*]
[*]Department of Medicine and Health Sciences “Vincenzo Tiberio”, University of Molise, Campobasso, Italy
{francesco.mercaldo, francesco.tariello, giuseppe.vanoli}@unimol.it
[†]Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy
francesco.mercaldo@iit.cnr.it
[‡]Department of Biosciences and Territory, University of Molise, Pesche (IS), Italy
antonella.santone@unimol.it

***Abstract*—Solar radiation significantly affected the cooling requirements of air-conditioned buildings during the summer period, attention is also paid to it in order to optimize the management of indoor lighting as it is a natural lighting source. However, the solar radiation is measured by a few weather stations in a few locations. The aim of this paper is to reconstruct the hourly solar global radiation trend on a year in some cities of the Northern and Southern Italy, starting from typical recorded meteorological data: temperature, relative humidity, wind speed and direction, etc. For this task a supervised machine learning algorithm has been used to build a model. The reached results show that the solar radiation hourly values can be extrapolated from other weather data in a reliable way, in fact an f-measure ranging from 0.950 an 1 is obtained for the several Italian cities involved in the experiment.**

I. Introduction and Related Work

Solar global radiation on a horizontal plane at the Earth surface on one hand depends on deterministic variables like: latitude of the place, the hour of the day, the day of the year, while, on the other hand it is significantly affected by stochastic parameters like the atmospheric conditions (cloudiness conditions). The measurement of the solar radiation all over the world is not so diffuse as other climatic quantities are. Weather stations often do not provide the measurement of solar radiation. Statistically in the world only one station every 500 stations records solar radiation [32].

As reported by the World Meteorological Organization in [1], solar radiation measurement is mandatory only in weather stations used for climate applications, it is not required for basic on land and marine meteorological observations, or for aeronautical meteorology. Therefore, data from aeronautics, can be uncomplete. The climatic stations installed in the cities have the aim to monitor the urban microclimate and usually do not provide the radiation data because they are located within an urbanized space and suffer from the shading of the buildings. A further issue that can take place concerns the recorded data, in some cases even where all data are measured, only maximum, minimum and average values of each day are then archived. Finally, the sensors for solar radiation are more expensive than the other and so are not provided if not necessary.

This determines that there are not many data to analyze and use in research activities. Solar radiation is the primary driver for the life on the Earth, it is at the base of many chemical, physical and biological processes. It plays a big role in the energy balance of the building. As concern this last aspect it can be very useful to know the solar radiation values because it could constitute the most significant energy contribution in the energy balance of a building during the summer period. Solar radiation knowledge is also important in the natural/artificial lighting management or in the air-conditioning plants design, sizing and operation especially if they are solar-activated. The energy audits of the buildings provide for an evaluation phase of the building-plant system improvement interventions. These assessments are carried out through simulation models. To demonstrate their reliability, it is necessary to validate and calibrate the models with respect to the measured environmental conditions and the energy demands that are, for example, obtained from the bills. The ISO 52016 of 2017 is the reference standard for the assessment of Energy performance of buildings, heating and cooling needs, internal temperatures and sensible and latent heat loads. It suggests an hourly method as the most detailed calculation solution. As mentioned before, the contribution of solar radiation is very important in the energy balance of a building and since bills refer to previous years and the detailed methodology considers hourly simulation, the importance of reconstructing hourly solar radiation for those places where it is not measured or not archived, appears evident. Many studies are focused on the solar radiation trend reconstruction, those estimations are performed on different time scales depending on their application: conventionally in meteorology the analyses are on annual base because they are used for the evaluation of climate change. On the contrary, a detailed approach (i.e., with a lower time resolution) is applied in the forecast activity, especially if it is for the estimation of the power production by renewable energy based plant connected to an electric grid. In [35] a sub-hourly time-step and a statistical approach are adopted in the solar radiation determination. Authors obtain a Root Mean Square Error (RMSE) value achievable are 23.4% and 7.2% on average. The models used to replicate the solar energy arriving on

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

the Earth can be classified as linear or non-linear, they are based on Angström-Prescott model [29] or on its modified versions. The time scale at which they have been applied was the day, the month and the year. More recent solutions, instead, consider artificial neural networks and fuzzy logic techniques [17], with an RMSE ranging from 2.01% to 9.32%. Chiteka and Enweremadu [9] designed a Neural Network to predict the global solar radiation in Zimbabwe. The geographical data of latitude and longitude and meteorological data of humidity pressure, clearness index and average temperature were used as inputs variables of the model. They exploit a network with 10 neurons with a tansig transfer function. The network achieved a determination coefficient equal to 99.89%. Researchers in [18] propose two networks to predict the solar irradiance by reporting a correlation coefficient more than 0.95 with regard to the testing data-set. As demonstrated by current literature [11], machine learning models outcome empirical models for predicting daily global solar radiation from sunshine duration.

The main difference with respect to these papers is represented by the reconstruction of the solar radiation at hourly rate with respect to a long period, a year. Using data-sets of two parts of Italy, i.e., from the Lombardia (in the North of Italy) and the Puglia region (in the South of Italy), a supervised machine learning model has been trained and then evaluated on several Italian cities. The final aim is to demonstrate the model ability to generalise different Italian cities from a climatic point of view. Unfortunately, other data, for instance related to more places and with a detailed time resolution, are not available but in future works the method will be extended. Despite being a preliminary work, this paper shows some novelty notes with respect to other works found in current literature: on one hand the reconstruction presented hereinafter has a certain degree of generality in terms of time, as it is performed for a long period (1 year) and with a short time-step (1 hour), succeeding in the determination of the solar radiation in all the four seasons of the year and not in one day or in some days. On the other hand, the proposed model is also appreciable in geographical terms for its capability to represent the radiation in cities that have different climatic conditions, as they are in the norther and southern part of Italy, reaching in each case very high levels of precision. The current state of art provides that software used for energy audit and dynamic simulations typically refer to weather data derived from historical series recorded in the selected location and/or from interpolations of data measured in more or less close climatic stations [17]. Typically, for many locations measured data like temperature, humidity, wind speed and direction are available but can not be used because information on the radiation is missing. Therefore, the aim of this paper is to reconstruct global solar radiation on horizontal surface using standard weather data in order to make the results of the aforementioned energy assessments more reliable. Furthermore, the proposed approach can be integrated in weather station software to estimate solar radiation also where it is not measured.

The remaining of the paper proceeds as follows: in the next section the proposed method aimed to reconstruct solar radiation is presented, in Section III an experiment aimed to demonstrate the effectiveness of the proposed method is provided and, finally, in the last section conclusion remarks and future works are drawn.

II. The Method

In this section we describe the proposed method to reconstruct the average solar radiation values starting from meteorological data.

To this aim, machine learning is considered. Machine learning is a type of artificial intelligence aimed to provide computers with the ability to learn without being explicitly programmed [28].

Machine learning tasks are typically classified into two categories, depending on the nature of the learning available to a learning system:

- *Supervised learning*: the computer is presented with example inputs and their desired outputs, given by a “teacher”, and the goal is to learn a general rule that maps inputs to outputs. It represents the classification: the process of building a model of classes from a set of records that contains class labels;
- *Unsupervised learning*: no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

In this paper we experiment the effectiveness of supervised classification to reconstruct solar radiation value. In fact, the most used algorithms for the model building tasks are typically supervised decision tree-based i.e., algorithms using a decision tree as a model which maps observations about an item (represented in the branches) to conclusions about the target (i.e., the label to reconstruct, the average solar radiation in the following study) of the items value (represented in the leaves). These algorithms (for instance, *LadTree*, *J48*, *RandomTree*, *RepTree*) are the most widespread to solve data mining problems [28] for instance, from malware detection [3, 26, 7, 22] to disease classification [24].

In detail the proposed method, coherently with the supervised machine learning classification, is mainly composed by two phases: the *training* (shown in Figure 1) and the *testing* (shown in Figure 2).

Figure 1 depicts the *training* phase of the proposed method.

The aim of this step is to build a model with the ability to reconstruct the solar radiation. The inputs of the *training* step are the following: a set of *Meteorological Data*, the relative average *Solar Radiation* (in fact, as previously explained, the supervised machine learning model training task requires the label to generate the model) and the *Time* intervals. In detail the values of the *Solar Radiation* and the *Meteorological Data* are gathered each hour, therefore we consider hourly measurements.

Table I shows the *Meteorological Data* we considered.

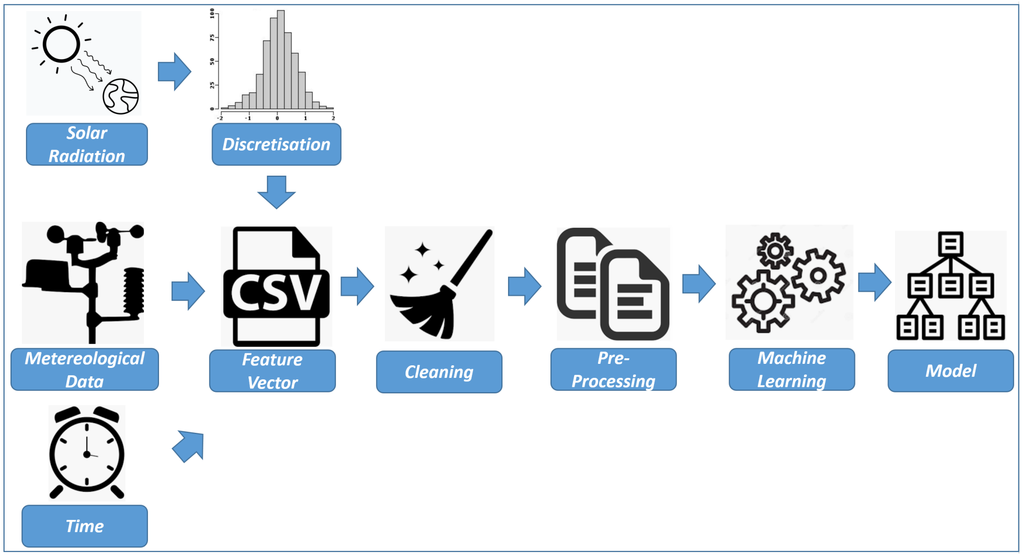

Fig. 1: The *training* step.

Feature	*Description*	*Measure unit*
F_1	Average hourly temperature	$^{\circ}$C
F_2	Maximum hourly temperature	$^{\circ}$C
F_3	Average hourly humidity	%
F_4	Minimum hourly humidity	%
F_5	Maximum hourly humidity	%
F_6	Prevailing wind direction	$^{\circ}$
F_7	Average wind speed	m/s
F_8	Maximum wind speed	m/s
F_9	Average atmospheric pressure	hPa
F_{10}	Accumulated precipitation	mm
F_{11}	Degree day	number of days

TABLE I: The *Meteorological Data*.

The features from F_1 to F_10 are practically available from the weather stations, from amateur weather stations to professional ones [11].

The F_{11} feature, the degree day, is a measure of heating or cooling. This measure is usually considered within an energy monitoring and targeting scheme to monitor the heating and cooling costs of climate controlled buildings, while annual figures can be used for estimating future costs.

The F_{11} feature is obtained as follows, according to UNI EN ISO 15927-6: 2008[1] i.e., the official Italian version of the European standard EN ISO 15927-6[2]:

$$F_{11} = \sum_{e=1}^{n} (T_0 - T_e)$$

where n represents the number of days of the conventional heating period, T_0 is conventional temperature and T_e average daily outdoor temperature such that $T_e < T_0$. For the definition, it is clear that the values related to F_{12} feature are different for places at different latitude and longitude: we refer to the official documentation[3] to obtain the degree day for the several localities analysed in the experimental analysis.

As previously explained, the supervised classification task requires the label to built an efficient model. For this reason the variable to reconstruct i.e., the average *Solar Radiation* (measured in $[W/m^2]$) is considered. We perform a *Discretisation* on the numeric values of the average *Solar Radiation*. In fact, data are usually given in the form of continuous values, the problem is that their number is huge and model building for such data can be difficult. Furthermore, many supervised machine learning algorithms operate only in discrete search or variable space [28]. For instance, decision trees typically divide the values of a variable into two parts according to an appropriate threshold value [27]. The aim of *Discretisation* is to reduce the number of values of a continuous variable assumes by grouping them into a number b of classes. Several techniques were proposed by research community to discretise continuous variables in discrete ones, in this paper we resort to the *equal-frequency discretization* [19]. Basically, this algorithm determines the minimum and maximum values of the numeric variable to discretise (in the following study the average solar radiation), sorts all values in ascending order, and divides the range into a user-defined number of intervals (we set $b = 10$), in such a way that every interval contains the equal number of sorted values. Table II shows the discretisation classes gathered by the discretisation technique.

As shown in Table II there are the 10 intervals in which the average solar radiation can falls, coherently with the number of bins $b = 10$ we set.

Once obtained the discretised values, we put together the discretised *solar radiation*, the *meteorological data* with the

[1] http://efficienzaenergetica.acs.enea.it/doc/dpr412-93_allA_tabellagradigiorno.pdf

[2] https://www.degreedays.net/#generate

[3] http://efficienzaenergetica.acs.enea.it/doc/dpr412-93_allA_tabellagradigiorno.pdf

Class	*Solar radiation range*
1	$(-\infty$ - 86.9]
2	(86.9 - 173.8]
3	(173.8 - 260.7]
4	(260.7 - 347.6]
5	(347.6 - 434.5]
6	(434.5 - 521.4]
7	(521.4 - 608.3]
8	(608.3 - 695.2]
9	(695.2 - 782.1]
10	(782.1 - ∞)

TABLE II: the *Discretisation* classes (ranges are expressed in $[W/m^2]$).

relative hourly *time* to generate the *Feature Vector* to train the model.

The next step is the feature vector *Cleaning*: it is aimed to remove all the instances where the *Solar Radiation* is equal to 0: in fact we are interested to reconstruct the radiation values different from zero, for this reason we consider not significant the instances where the solar radiation value is equal to 0 [15]. Moreover, considering that during the night hours the radiation is always equal to 0, this would lead to an introduction of noise in the data which would lead to inaccuracies in the model.

After the *Cleaning* step, the next one is represented by the *Pre-Processing*: this step is focused on removing equal instances in the feature vector. It is advisable to remove duplicates while segregating samples for training and testing [33]. We basically remove duplicate instances such that only one of all the duplicate instances is kept. Two (or more) instances are considered duplicate if the *meteorological data*, the *time* and the *solar radiation* exhibit the same values in the considered instances.

Then, in the *Machine Learning* step we adopt a supervised machine learning classification algorithm to generate a *Model*. In detail, in this paper we experiment the effectiveness of solar radiation reconstruction of a decision tree algorithm, the *LadTree* [14] one. We exploit this algorithm considering its ability to generate a multi-class alternating decision tree using the LogitBoost [16] strategy and its successfully application in different contexts, from the medical field [24] to the software security one [6, 5, 10]. In fact, considering that each instance under analysis must be assigned in one of the *b* solar radiation classes, this represents a multi-class classification problem. In machine learning, multiclass or multinomial classification is the problem of classifying instances into one of three or more classes (in our case into one of ten instances). While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by using several strategies [28].

Once generated the *Model*, the *testing* phase is focused on the analysis of its accuracy for solar radiation reconstruction.

Figure 2 shows the *testing* phase related to the proposed method.

The *meteorological data* and the *time* step are the same we described for the *training phase*: the only difference is that in this case the average *solar radiation* is the variable to reconstruct. Clearly, also the *cleaning* and the *Pre-Processing* steps are the same we explained in the *training* phase. In fact, the *feature vector* is cleaned by all the instances where the average solar radiation is equal to 0 and from instance duplicates.

Thus, the *feature vector* is an input for the *Model* that will reconstruct one of the 10 solar radiation intervals for each instance under analysis.

III. The Experimental Analysis

The experimental analysis section describes the experiment we designed to evaluate the effectiveness of the proposed method in average solar radiation.

A. The Data-set

Meteorological data belonging to several localities (i.e., Milano, Bari, Cremona, Mantova, Brindisi and Benevento) in Italy were obtained from different sources. The localities involved are shown in Figure 3.

As shown from Figure 3, localities in different latitude and longitude are chosen, to demonstrate the effectiveness of the proposed method to build a classifier able to generalise the global solar radiation, regardless of the specific locality.

In Table III we report for each considered locality the altitude, the longitude and the degree day (the last one according with the information provided by the UNI EN ISO 15927-6: 2008 document[4].

Locality	*Latitude*	*Latitude*	*Degree day*
Milano	45.46	9.18	2404
Bari	41.12	16.86	1185
Benevento	41.12	14.78	1316
Cremona	45.13	10.02	2389
Mantova	45.15	10.79	2388
Brindisi	40.63	17.94	1083

TABLE III: The *degree days* for the localities involved in the experiment (expressed in number of days).

The meteorological data for the cities of Bari and Brindisi were obtained from “ARPA Puglia” (Agenzia Regionale per la Prevenzione e la Protezione Ambientale)[5], the Regional Agency for Environmental Protection of Puglia. The data for the Bari and Brindisi cities are available at the following url: http://www.arpa.puglia.it/web/guest/serviziometeo.

The data for the Milano, Mantova and Cremona localities were gathered from the “ARPA Lombardia” (Agenzia Regionale per la Protezione dell’Ambiente della Lombardia)[6], the Regional Agency for the Environmental Protection of Lombardia. The data for Milano, Mantova and Cremona are available at the following url:

[4] http://efficienzaenergetica.acs.enea.it/doc/dpr412-93_allA_tabellagradigiorno.pdf

[5] http://www.arpa.puglia.it/web/guest/serviziometeo

[6] https://www.arpalombardia.it/Pages/ARPA_Home_Page.aspx

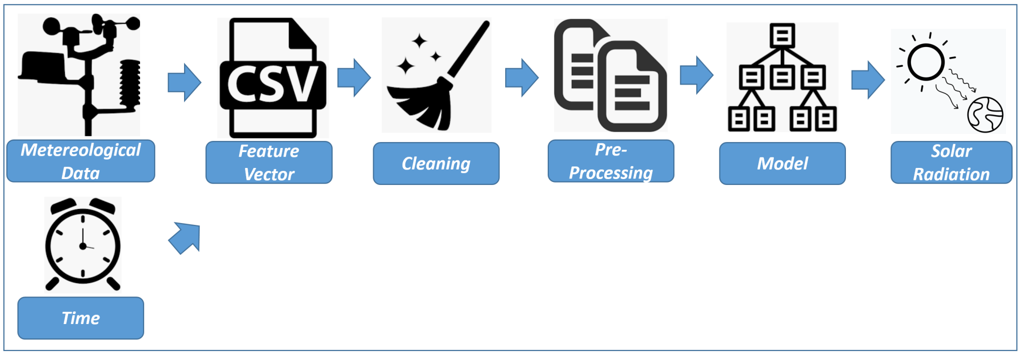

Fig. 2: The *testing* step.

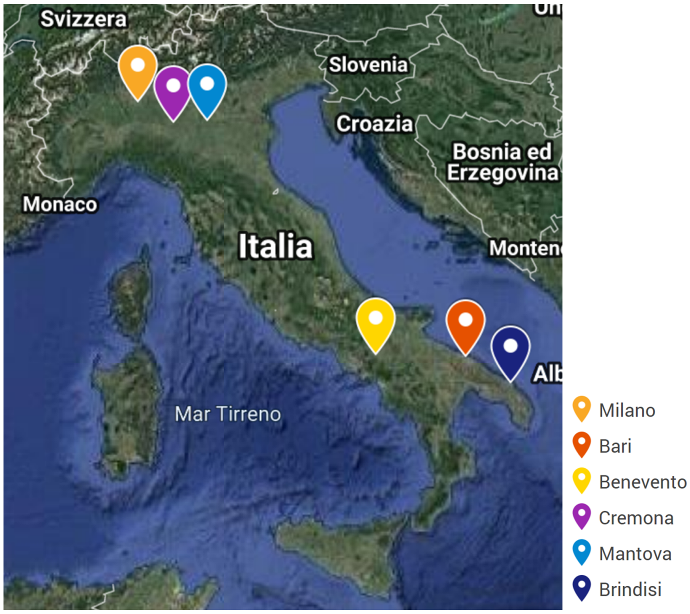

Fig. 3: Map of Italy with the localities involved in the experiment.

https://www.arpalombardia.it/siti/arpalombardia/meteo/richiesta-dati-misurati/Pagine/RichiestaDatiMisurati.aspx by submitting a web form.

The data for the city of Benevento were gathered from a weather station provided by the University of Sannio[7] in Benevento.

B. The Classification

The classification analysis consisted of building several classifiers to evaluate the meteorological data accuracy to discriminate between global solar radiation in different localities.

Figure 4 shows the overall picture of the experiment settings.

As shown in Figure 4, the cities of Milano and Bari are selected for the model building phase (depicted in Figure 1), while for the testing phase we considered the remaining cities (depicted in Figure 2): Cremona, Mantova, Brindisi and Benevento.

We considered the cities of Milano and Bari for the model training considering that from a climatic point of view they can be considered the antipodes [21]. In fact, the Bari climate is typically Mediterranean, with mild winters and hot, long and very often humid summers[8]. As a seaside city Bari, thanks to the mitigating action of the Adriatic sea from which it is wet, it has a more typically maritime climate, with less pronounced seasonal temperature ranges [20]. Differently, Milano has a continental climate, characterized by hot and humid summers and harsh ones and in the winter months moderate precipitation occurs, but also some snowfall [9].

Considering the proximity of Brindisi to Bari, we can state that from a climatic point of view the two cities are similar, for the same reason we consider the cities of Cremona and Mantova close to Milano (as appears in Figure 3).

The rationale behind the adoption of two the cities of Milano and Bari for the model building is to create a model as general as possible. In machine learning, generalization usually refers to the ability of an algorithm to be effective across a range of inputs: to demonstrate the generalisation of the model we perform the evaluation on others cities, several really closer to Milano (i.e., Cremona and Mantova), to Bari (i.e., Brindisi) and Benevento (similar neither in Milan nor in Bari). In fact in Benevento there is a warm and temperate climate. There is more rainfall in the winter than in the summer in Benevento. The climate of Benevento has more continental stretches than the maritime one of the areas of Caserta and Napoli. In the winter semester the temperature is generally lower; the rains are relatively frequent, as well as fog, frost, and sometimes frost (with temperatures of some degree below zero).

For training the classifiers, we defined T as a set of labeled messages *(M, l)*, where each M is the label associated to the global solar radiation ranges $l \in \{$ *(-∞ - 86.9], (86.9 - 173.8], (173.8 - 260.7], (260.7 - 347.6], (347.6 - 434.5], (434.5 - 521.4], (521.4 - 608.3], (608.3 - 695.2], (695.2 - 782.1], (782.1 -* ∞*)*$\}$. For each M we built a feature vector $F \in R_y$,

[7] https://www.unisannio.it/

[8] https://www.ilmeteo.it/

[9] https://www.accuweather.com/it/it/italy-weather

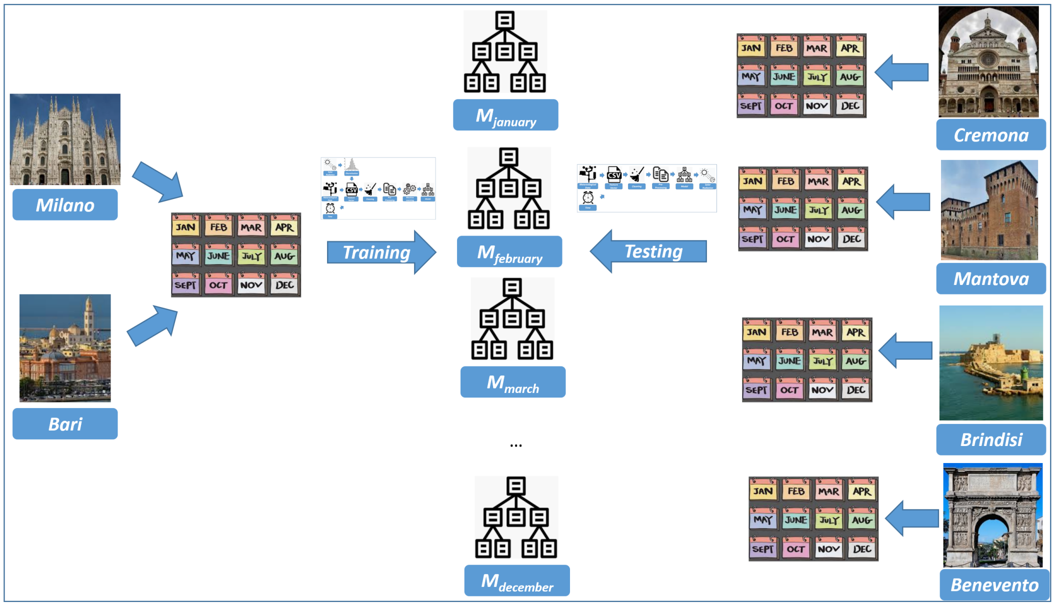

Fig. 4: The experiment settings.

where y is the number of the features used in training phase ($y = 11$).

For the learning phase, we use a k-fold cross-validation: the data-set is randomly partitioned into k subsets. A single subset is retained as the validation data-set for testing the model, while the remaining $k - 1$ subsets of the original data-set are used as training data. We repeated the process for $k = 10$ times; each one of the k subsets has been used once as the validation data-set. To obtain a single estimate, we computed the average of the k results from the folds.

We evaluated the effectiveness of the classification method for the cities of Milano and Bari with the following procedure:

1) build a training set $T \subset D$;
2) build a testing set $T' = D \div$T;
3) run the training phase on T;
4) apply the learned classifier to each element of T'.

Each classification was performed using 80% of the data-set as training data-set and 20% as testing data-set employing the full feature set exploiting the *LadTree* [28] classification algorithm.

Moreover, the data related to the cities of Cremona, Mantova, Benevento and Brindisi are considered as testing set.

Furthermore, for the cities involved in the experiment i.e., Milano, Bari, Benevento, Cremona, Mantova and Brindisi the values related to the meteorological data and the solar global radiation were obtained for several years. In fact, data from the following years were available from the considered data sources, as shown in Table IV.

In Table IV, the ✓symbol indicates a year considered to train the model, the ✓ stands for the a year available in the data-sets and considered to evaluated the model, while the the ✗ symbol denotes the non availability of the year.

We generate, as shown in Figure 4, a model for each month (we built a total of 12 models).

Thus, considering that we obtained data for several years, according to the Table IV we obtained several sub L^y_m data-sets where $L \in \{$Milano, Bari, Cremona, Mantova, Benevento, Brindisi$\}$, $m \in \{$January, February, March, April, May, June, July, August, September, October, November, December$\}$ and $y \in \{2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018\}$.

According to Table IV, we build a model for each m month: below we indicate the L^y_m data-sets considered to train each M_m model where $m \in \{$January, February, March, April, May, June, July, August, September, October, November, December$\}$.

The twelve M_m models are evaluated with the cross-validation, while the remaining sub data-sets of Milano and Bari (i.e., $Milano^{2018}_m$, $Bari^{2015}_m$, $Bari^{2016}_m$, $Bari^{2017}_m$ and $Bari^{2018}_m$) are considered as testing data-sets for the trained models. Clearly to evaluate, for instance, $M_{december}$ we considered sub $L^y_{december}$ data-sets (i.e., sub data-sets related to the same month of the trained model).

The following metrics are considered to evaluate the classification results [28, 27]: Precision, Recall and F-Measure.

Table V shows the average performances we obtained.

As appears in Table V, the precision is ranging from 0,986 for the (173.8 - 260.7] solar radiation class to 1,000 for the (-∞ - 86.9] class. The lowest recall value reached is 0,905 for (173.8 - 260.7] solar radiation class, while the remaining classes obtain a recall equal to 1,000.

The experimental results demonstrate the effectiveness of the proposed that for global solar radiation reconstruction.

IV. Conclusion and Future Work

Considering the importance of solar radiation measurements and that only few weather station provides this parameter, in

Locality	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
Milano	✗	✗	✗	✗	✗	✓	✗	✓	✓
Bari	✓	✓	✓	✓	✓	✓	✓	✓	✓
Benevento	✗	✗	✗	✗	✗	✗	✓	✓	✗
Cremona	✗	✗	✗	✗	✗	✓	✓	✓	✓
Mantova	✗	✗	✗	✗	✗	✓	✓	✓	✓
Brindisi	✓	✓	✓	✓	✓	✓	✓	✓	✓

TABLE IV: The years available for each locality, with details about the years used in the model training and testing.

TABLE V: Performance evaluation.

Precision	*Recall*	*F-Measure*	*Class*
1,000	0,905	0,950	$(-\infty$ - $86.9]$
0,980	1,000	0,990	(86.9 - 173.8]
0,988	1,000	0,994	(173.8 - 260.7]
0,986	1,000	0,993	(260.7 - 347.6]
0,990	1,000	0,995	(347.6 - 434.5]
0,988	1,000	0,994	(434.5 - 521.4]
0,990	1,000	0,995	(521.4 - 608.3]
0,991	1,000	0,996	(608.3 - 695.2]
0,994	1,000	0,997	(695.2 - 782.1]
0,999	1,000	1,000	$(782.1$ - $\infty)$

this paper a method to reconstruct the solar radiation starting from a set of variables currently measured also from cheaper weather station is proposed. Supervised machine learning is exploited, with the aim to build a model able to reconstruct the solar radiation for different Italian cities. Experiments on real data gathered from several reliable sources demonstrated the effectiveness of the proposed method by reaching an f-measure ranging from 0.950 to 1. Future research lines will consider the adoption of deep learning [25, 4, 23, 34] and formal verification techniques [8, 12, 30, 2, 31, 13] with the aim to improve the performances on an extended data-set and to reconstruct the exact values related to global solar radiation (not the discretisation range). Another research direction is related to the adoption of unsupervised approaches i.e., the application of algorithms for generating clusters of data without the radiation label.

ACKNOWLEDGEMENT

The authors are thankful to ARPA *U.O. Servizio Metereologico e Rete Idrometeo regionale*[10] for providing us the data related to the Lombardia region (in particular for the cities of Milano, Cremona and Mantova).

REFERENCES

[1] World Meteorological Organization, Manual on the WMO Integrated Global Observing System, Annex VIII to the WMO Technical Regulations, WMO-No. 1160 . https://library.wmo.int/index.php?lvl=notice_display&id=19223#.XpWKz8gzZPY, 2019. Online; accessed 14 March 2020.

[2] Roberto Barbuti, Nicoletta De Francesco, Antonella Santone, and Gigliola Vaglini. Reduced models for efficient ccs verification. *Formal Methods in System Design*, 26(3):319–350, 2005.

[3] Pasquale Battista, Francesco Mercaldo, Vittoria Nardone, Antonella Santone, and Corrado Aaron Visaggio. Identification of android malware families with model checking. In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy, ICISSP 2016, Rome, Italy, February 19-21, 2016.*, pages 542–547. SciTePress, 2016.

[4] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. An ensemble learning approach for brain cancer detection exploiting radiomic features. *Computer methods and programs in biomedicine*, 185:105134, 2020.

[5] Gerardo Canfora, Francesco Mercaldo, and Corrado Aaron Visaggio. Mobile malware detection using op-code frequency histograms. In *2015 12th International Joint Conference on e-Business and Telecommunications (ICETE)*, volume 4, pages 27–38. IEEE, 2015.

[6] Gerardo Canfora, Francesco Mercaldo, and Corrado Aaron Visaggio. An hmm and structural entropy based detector for android malware: An empirical study. *Computers & Security*, 61:1–18, 2016.

[7] Gerardo Canfora, Francesco Mercaldo, Corrado Aaron Visaggio, and Paolo Di Notte. Metamorphic malware detection using code metrics. *Information Security Journal: A Global Perspective*, 23(3):57–67, 2014.

[8] Michele Ceccarelli, Luigi Cerulo, and Antonella Santone. De novo reconstruction of gene regulatory networks from time series data, an approach based on formal methods. *Methods*, 69(3):298–305, 2014.

[9] K Chiteka and CC Enweremadu. Prediction of global horizontal solar irradiance in zimbabwe using artificial neural networks. *Journal of Cleaner Production*, 135:701–711, 2016.

[10] Mario GCA Cimino, Nicoletta De Francesco, Francesco Mercaldo, Antonella Santone, and Gigliola Vaglini. Model checking for malicious family detection and phylogenetic analysis in mobile environment. *Computers & Security*, 90:101691, 2020.

[11] Junliang Fan, Lifeng Wu, Fucang Zhang, Huanjie Cai, Wenzhi Zeng, Xiukang Wang, and Haiyang Zou. Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review

[10] https://www.arpalombardia.it/Pages/ARPA_Home_Page.aspx

and case study in china. *Renewable and Sustainable Energy Reviews*, 100:186–212, 2019.

[12] Nicoletta de Francesco, Giuseppe Lettieri, Antonella Santone, and Gigliola Vaglini. Grease: a tool for efficient nonequivalence checking. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 23(3):24, 2014.

[13] Sara Gradara, Antonella Santone, and Maria Luisa Villani. Using heuristic search for finding deadlocks in concurrent systems. *Information and Computation*, 202(2):191–226, 2005.

[14] Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, Eibe Frank, and Mark Hall. Multiclass alternating decision trees. In *European Conference on Machine Learning*, pages 161–172. Springer, 2002.

[15] Muhammad Iqbal. *An introduction to solar radiation*. Elsevier, 2012.

[16] Sushilkumar Rameshpant Kalmegh. Comparative analysis of weka data mining algorithm randomforest, randomtree and ladtree for classification of indigenous news data. *International Journal of Emerging Technology and Advanced Engineering*, 5(1):507–517, 2015.

[17] Tamer Khatib, Azah Mohamed, and Kamaruzzaman Sopian. A review of solar energy modeling techniques. *Renewable and Sustainable Energy Reviews*, 16(5):2864–2869, 2012.

[18] A Khosravi, RNN Koury, L Machado, and JJG Pabon. Prediction of hourly solar radiation in abu musa island using machine learning algorithms. *Journal of Cleaner Production*, 176:63–75, 2018.

[19] Sotiris Kotsiantis and Dimitris Kanellopoulos. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

[20] David Maddison and Andrea Bigano. The amenity value of the italian climate. *Journal of Environmental Economics and Management*, 45(2):319–332, 2003.

[21] C Marino, A Nucara, and M Pietrafesa. Does window-to-wall ratio have a significant effect on the energy consumption of buildings? a parametric analysis in italian climate conditions. *Journal of Building Engineering*, 13:169–183, 2017.

[22] Fabio Martinelli, Fiammetta Marulli, and Francesco Mercaldo. Evaluating convolutional neural network for effective mobile malware detection. *Procedia Computer Science*, 112:2372–2381, 2017.

[23] Fabio Martinelli, Francesco Mercaldo, Vittoria Nardone, and Antonella Santone. Car hacking identification through fuzzy logic algorithms. In *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on*, pages 1–7. IEEE, 2017.

[24] Francesco Mercaldo, Vittoria Nardone, and Antonella Santone. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. *Procedia Computer Science*, 112(C):2519–2528, 2017.

[25] Francesco Mercaldo and Antonella Santone. Deep learning for image-based mobile malware detection. *Journal of Computer Virology and Hacking Techniques*, pages 1–15.

[26] Francesco Mercaldo, Corrado Aaron Visaggio, Gerardo Canfora, and Aniello Cimitile. Mobile malware detection in the real world. In *Software Engineering Companion (ICSE-C), IEEE/ACM International Conference on*, pages 744–746. IEEE, 2016.

[27] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[28] Tom M Mitchell. Machine learning and data mining. *Communications of the ACM*, 42(11):30–36, 1999.

[29] M Paulescu, N Stefu, D Calinoiu, E Paulescu, N Pop, R Boata, and O Mares. Ångström–prescott equation: Physical basis, empirical models and sensitivity analysis. *Renewable and Sustainable Energy Reviews*, 62:495–506, 2016.

[30] Antonella Santone. Automatic verification of concurrent systems using a formula-based compositional approach. *Acta Informatica*, 38(8):531–564, 2002.

[31] Antonella Santone. Clone detection through process algebras and java bytecode. In *IWSC*, pages 73–74. Citeseer, 2011.

[32] Huaiwei Sun, Na Zhao, Xiaofan Zeng, and Dong Yan. Study of solar radiation prediction and modeling of relationships between solar radiation and meteorological variables. *Energy Conversion and Management*, 105:880–890, 2015.

[33] Xiaoyuan Xie, Joshua WK Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011.

[34] Hanqi Zhang, Xi Xiao, Francesco Mercaldo, Shiguang Ni, Fabio Martinelli, and Arun Kumar Sangaiah. Classification of ransomware families with machine learning based on n-gram of opcodes. *Future Generation Computer Systems*, 90:211–221, 2019.

[35] Xiongwen Zhang. A statistical approach for sub-hourly solar radiation reconstruction. *Renewable energy*, 71:307–314, 2014.